

# The American Economic Review

Cuk-1402184-81-P007830

81

## ARTICLES

- G. ACKLEY      **Commodities and Capital: Prices and Quantities**
- M. FELDSTEIN AND J. GREEN      **Why Do Companies Pay Dividends?**
- E. E. LEAMER      **Let's Take the Con Out of Econometrics**
- J. A. WILCOX      **Why Real Interest Rates Were So Low in the 1970's**
- G. A. AKERLOF      **Loyalty Filters**
- L. DE ALESSI      **Property Rights, Transaction Costs, and X-Efficiency**
- F. M. FISHER AND J. J. MCGOWAN      **On the Misuse of Accounting Rates of Return to Infer Monopoly Profits**
- J. V. HENDERSON AND Y. M. IOANNIDES      **A Model of Housing Tenure Choice**
- J. MAYSHAR      **On Divergence of Opinion and Imperfections in Capital Markets**
- W. J. FRAZER, JR. AND L. A. BOLAND      **An Essay on the Foundations of Friedman's Methodology**
- C. ASH, B. UDIS, AND R. F. MCNOWN      **Enlistments in the All-Volunteer Force**
- D. R. SHALLER      **Working Capital Finance Considerations in National Income Theory**
- S. A. WOODBURY      **Substitution Between Wage and Nonwage Benefits**

SHORTER PAPERS: J. Eaton and S. J. Turnovsky; H. P. Marvel and E. J. Ray; F. Cesarano; W. R. Johnson and E. K. Browning; E. Koskela and M. Virén; D. L. Coursey and V. L. Smith; J. B. Burbidge; A. B. Abel; B. L. Copeland, Jr.; R. J. Shiller; C. Daniel III; T. F. Bresnahan; S. Yabushita; J. Stiglitz and A. Weiss.

MARCH 1983

# THE AMERICAN ECONOMIC ASSOCIATION

●Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Dues for 1983, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:

\$32.00 for regular members with rank of assistant professor or lower, or with annual income of \$15,350, or less;

\$38.40 for regular members with rank of associate professor, or with annual income of \$15,350 to \$25,600;

\$44.80 for regular members with rank of full professor, or with annual income above \$25,600;

\$16.00 for junior members (registered students). Certification must be submitted yearly.

Subscriptions (libraries, institutions, or firms) are \$100.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$9.20 to cover extra postage.

●Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212.

Founded in 1885

## Officers

### *President*

W. ARTHUR LEWIS  
Princeton University

### *President-Elect*

CHARLES L. SCHULTZE  
The Brookings Institution

### *Vice Presidents*

JUANITA M. KREPS  
Duke University  
EDMUND S. PHELPS  
Columbia University

### *Secretary*

C. ELTON HINSHAW  
Vanderbilt University

### *Treasurer*

RENDIGS FELS  
Vanderbilt University

### *Managing Editor of The American Economic Review*

ROBERT W. CLOWER  
University of California-Los Angeles

### *Managing Editor of The Journal of Economic Literature*

MOSES ABRAMOVITZ  
Stanford University

## Executive Committee

### *Elected Members of the Executive Committee*

ELIZABETH E. BAILEY  
Civil Aeronautics Board  
ROBERT J. GORDON  
Northwestern University  
ANN F. FRIEDLAENDER  
Massachusetts Institute of Technology  
JOSEPH E. STIGLITZ  
Princeton University  
WILLIAM D. NORDHAUS  
Yale University  
A. MICHAEL SPENCE  
Harvard University

### *EX OFFICIO Members*

WILLIAM J. BAUMOL  
Princeton University and New York University  
GARDNER ACKLEY  
The University of Michigan

# THE AMERICAN ECONOMIC REVIEW

ROBERT W. CLOWER  
Managing Editor

JOHN G. RILEY  
Associate Editor

WILMA ST. JOHN  
Production Editor

THERESA DE MARIA  
Assistant Editor

## Board of Editors

GEORGE A. AKERLOF  
ALBERT ANDO  
G. O. BIERWAG  
THOMAS F. COOLEY  
PATRICIA DANZON  
RONALD G. EHRENBERG  
H. E. FRECH III  
HERSCHEL I. GROSSMAN  
JACK HIRSHLEIFER  
PETER W. HOWITT  
LAURENCE J. KOTLIKOFF  
ANNE O. KRUEGER  
FREDERIC S. MISHKIN  
SHERWIN ROSEN  
RICHARD SCHMALENSEE  
JAMES P. SMITH  
E. ROY WEINTRAUB  
ROBERT D. WILLIG

● Manuscripts and editorial correspondence relating to the regular quarterly issues of this *Review* and the *Papers and Proceedings* should be addressed to Robert W. Clower, Managing Editor, *AER* Editorial Office, University of California, Los Angeles, CA 90024. Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: \$25 for members; \$50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

● No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

● Copyright © American Economic Association 1983. All rights reserved.

March 1983

VOLUME 73, NUMBER 1

## Articles

- Commodities and Capital: Prices and Quantities  
*Gardner Ackley* 1
- Why Do Companies Pay Dividends?  
*Martin Feldstein and Jerry Green* 17
- Let's Take the Con Out of Econometrics  
*Edward E. Leamer* 31
- Why Real Interest Rates Were So Low in the 1970's  
*James A. Wilcox* 44
- Loyalty Filters  
*George A. Akerlof* 54
- Property Rights, Transaction Costs, and X-Efficiency: An Essay in Economic Theory  
*Louis De Alessi* 64
- On the Misuse of Accounting Rates of Return to Infer Monopoly Profits  
*Franklin M. Fisher and John J. McGowan* 82
- A Model of Housing Tenure Choice  
*J. V. Henderson and Y. M. Ioannides* 98
- On Divergence of Opinion and Imperfections in Capital Markets  
*Joram Mayshar* 114
- An Essay on the Foundations of Friedman's Methodology  
*William J. Frazer, Jr. and Lawrence A. Boland* 129
- Enlistments in the All-Volunteer Force: A Military Personnel Supply Model and Its Forecasts  
*Colin Ash, Bernard Udis, and Robert F. McNown* 145
- Working Capital Finance Considerations in National Income Theory  
*Douglas R. Shaller* 156
- Substitution Between Wage and Nonwage Benefits  
*Stephen A. Woodbury* 166

391  
001

## Shorter Papers

Exchange Risk, Political Risk, and Macroeconomic Equilibrium	<i>Jonathan Eaton and Stephen J. Turnovsky</i>	183
The Kennedy Round: Evidence on the Regulation of International Trade in the United States	<i>Howard P. Marvel and Edward J. Ray</i>	190
The Rational Expectations Hypothesis in Retrospect	<i>Filippo Cesarano</i>	198
The Distributional and Efficiency Effects of Increasing the Minimum Wage: A Simulation	<i>William R. Johnson and Edgar K. Browning</i>	204
Social Security and Household Saving in an International Cross Section	<i>Erkki Koskela and Matti Virén</i>	212
Price Controls in a Posted Offer Market	<i>Don L. Coursey and Vernon L. Smith</i>	218
Government Debt in an Overlapping-Generations Model with Bequests and Gifts	<i>John B. Burbidge</i>	222
Optimal Investment under Uncertainty	<i>Andrew B. Abel</i>	228
Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?: Comment	<i>Basil L. Copeland, Jr.</i>	234
Reply	<i>Robert J. Shiller</i>	236
Duopoly Models with Consistent Conjectures: Comment	<i>Coldwell Daniel III</i>	238
Reply	<i>Timothy F. Bresnahan</i>	240
Theory of Screening and the Behavior of the Firm: Comment	<i>Shiro Yabushita</i>	242
Alternative Approaches to Analyzing Markets with Asymmetric Information: Reply	<i>J. Stiglitz and A. Weiss</i>	246
<b>Errata:</b>		
Product Differentiation Advantages of Pioneering Brands		250
Inventory Investment and the Theory of the Firm		251
Notes		252

$$\begin{array}{r} 391 \\ \hline 001 \end{array}$$

P 7830



*Gardner Ackley*

# Commodities and Capital: Prices and Quantities

By GARDNER ACKLEY\*

When I began to study economics, in the 1930's, macroeconomics certainly existed, in the works of such luminaries as Wicksell, Fisher, Robertson, and the early Keynes. But it was surely not recognized as a branch of economics in which one might specialize; nor was it regarded as necessary for the education of an economist. Thus, like almost everyone else, I began professional life as a price-theorist. The principal alternative was to become an institutionalist; and that didn't particularly attract me. Only later, and somewhat by accident, did I become a macroeconomist.

Ever since the emergence of macroeconomics as a distinct and (almost) respectable branch of analysis, there has been a conscious tension between macroeconomics and microeconomics; much of this tension relates to the roles and the behavior of prices and of the price level. Surely, it cannot be said that macroeconomics *ignores* prices and price changes, as is sometimes suggested. Milton Friedman and his monetarist associates and followers—who are macroeconomists of the first water—surely do not ignore prices. And the patron saint of *my* kind of macroeconomics, John Maynard Keynes, also certainly paid a great deal of attention to both relative prices and the price level. My concern with prices here, however, is not with inflation, but rather with the price-theoretical foundations of macroeconomics. Essentially, I will be discussing some of the roles that microeconomists and macroeconomists see for prices, and particularly for price changes.

Price theory has moved a long way since I deserted it, and I no longer claim any exper-

tise in this area. But I have the impression that many current problems both in micro- and macroeconomics tend to be the same problems—looked at from opposite sides of the borderline between them. Thus I propose to lead us on a stroll along some sectors of that border, moving back and forth across it from time to time, for there is no fence. And I intend only a meander, not a mapping. The spirit in which the journey is undertaken is that this is really all one country, and strollers should be welcome.

But it is not an imaginary country that I propose we visit. We will see no Walrasian auctioneers, although we will see many markets that seem to work pretty well without them. On the other hand, we will see very few wage rates being frequently revised; and many prices will look as though they were being revised only to maintain fairly stable markups over unit costs. In general, we will observe that the population of this country is neither very much brighter—nor much more stupid—than you and I are, in our own economic decisions.

Some aspects of the problems that I will discuss are particularly important in an age of inflation, and to the theory of inflation. But I prefer to conduct most of my discussion without explicit reference to changes in the general price level. For almost all of the matters that I will discuss primarily involve relative prices; and it is simpler to deal with them on the assumption of a constant price level. That way, I do not need to keep repeating “real price.”

I start with the prices and inventories of standardized commodities.

## I. Commodity Prices and Inventories

Economists often find it useful to think of the quantities of a commodity supplied to a competitive market as consisting of a supply from current production plus a possible supply from inventory; and the market de-

\*Presidential address delivered at the ninety-fifth meeting of the American Economic Association, December 29, 1982. I am most grateful to a number of colleagues and friends who read and commented on one or more earlier drafts of this paper, and especially to Saul H. Hymans, Hal Varian, John G. Cross, Theodore Bergstrom, Ronald Teigen, John Laitner, Charles L. Schultze, Otto Eckstein, and Henry Aaron.

mand, similarly, as consisting of a demand for current use or consumption plus a possible demand for additions to inventory. Any market-participant's demand for additions to inventory clearly should be based upon an expectation that the price will increase sufficiently over some future period, at least to cover costs of storage plus interest for that period. Any market-participant's willingness to supply goods from inventory should be based on an expectation that the price will rise insufficiently over any future period to cover costs of storage and interest for that period. Since expectations of future prices are never certain, the expected prices described should include discounts to reflect risk aversion, or premiums to reflect the pleasures of risk assumption.

Given the conditions that permit or require the holding of inventories, inventories may have either a stabilizing or a destabilizing effect on the commodity's price, production, and consumption. Traditionally, however, economists have assumed that the effect of inventories is to stabilize each of these. We think of the wheat market, for instance, where an entire year's new supply (at least in either Hemisphere) becomes available in a relatively short period of time. Were the holding of inventories of wheat (or wheat products) *impossible*, the market price would have to fall low enough during the harvest period to assure immediate consumption of the entire crop. The price would then rise high enough during the remainder of the year to cut off all consumption—or else high enough to make profitable the production of wheat in greenhouses, and to reduce its off-season consumption to the quantity so produced. However, given the possibility of *storage* of wheat or wheat products—by producers, dealers, or intermediate or final consumers—the price would normally fall only enough during the harvest period to induce the appropriate amount of inventory accumulation. The price would then rise, during the remainder of the year, as inventories were drawn down, sufficiently to cover at all times the costs of storage, interest, and the assumption of risk from holding inventories. Such seasonal price and inventory movements often are approximately observable.

The recognition that not only seasonal but also year-to-year variations may occur in the harvest as a result of weather conditions or civil disturbance, and that year-to-year variations may also occur in the demand for wheat, reflecting, say, the “business cycle,” will induce some or many market participants to carry over inventories of wheat (or of its products) from one crop year to the next (or to reduce stocks previously carried over). Stocks held will increase or decrease as the current market price varies relative to what is believed to be the current “normal” or “long-run equilibrium” price: the price that would exactly balance normal annual new supply and normal annual consumption. Through this process, not only seasonal fluctuations but also year-to-year fluctuations of market price and of consumption are reduced, even in the presence of substantial year-to-year changes in production or in final demand. Through changes in inventories, supply or demand in any particular year, in effect, may thus borrow from past and/or from future years' supply or demand.

As, and to the extent, that market participants learn of new “events” affecting or likely to affect either future normal production or future normal demand, the rate of inventory accumulation or decumulation will be varied so as to produce a new time path of price and current consumption, associated with an expected new equilibrium of prices and quantities. Needless to add, this behavior is stabilizing with respect both to prices and quantities: at least as compared to what would have happened if inventories could not exist, or if storage costs—or the interest rate—were higher. And this activity clearly increases economic welfare.

These activities of course reflect the behavior now described as based upon “rational expectations.” Market participants have in their heads a “model” of how supply and demand affect market price over time; on the basis of that model, and of all available information—including all *new* information—about present and future production and demand and the size of existing inventories, agents buy for or sell from inventories in ways that cause the market price to move continuously toward its current nor-

mal, long-term, market-clearing level, taking account of storage costs and interest, and appropriately discounted for uncertainty. Indeed, it was precisely in the context of reasoning about markets for storable individual commodities (and agricultural commodities in particular) that John Muth first presented the concept that he called "rational expectations".<sup>1</sup> In this particular microeconomic context there can surely be no valid objection to rational expectations. But we must not forget that the primary way in which expectations—rational or otherwise—affect current prices is precisely through affecting behavior with respect to inventories, along with sales or purchases for future delivery (i.e., negative inventories) or through affecting some other form of intertemporal substitution between, for example, labor and leisure, now and in the future.<sup>2</sup>

Markets for commodities, of course, vary widely in the extent to which inventories are able to stabilize price or consumption. At the one extreme is the traditional market for fresh flowers. Holding inventories is, by definition, impossible. Random fluctuations in the production of or the demand for fresh flowers are thus fully reflected in daily prices.

<sup>1</sup>Muth's 1961 article is, of course, now a classic. However, an excellent much-earlier statement of the way in which information and expectations affect the wheat price appears (of course!) in Alfred Marshall. The accompanying marginal note reads "Nearly all dealings in commodities that are not very perishable, are affected by calculation of the future" (pp. 337–38). Marshall perhaps failed to close his argument by stating explicitly that expected prices enter into today's production decisions, although that is clearly implied.

<sup>2</sup>It was in precisely the context of inventories—and by explicit analogy with the wheat market—that I once explained Keynes' theory of the "speculative demand" for money (or bonds). In my *Macroeconomic Theory* (pp. 175–76), I argued that, in the bond market—as in the wheat market—speculation on future bond prices, in the presence of large outstanding stocks of bonds and of money, by agents who form expectations of "normal" market-clearing prices, tends to stabilize bond prices at or close to their normal level—thereby preventing interest rate fluctuations from stabilizing aggregate demand for goods and services in the short run against exogenous variations in saving or investment propensities. Although the term rational expectations had not yet been invented, I was arguing that rational expectations explained an interest-elastic (or bond-price-elastic) demand for money.

And because the immediate response of amounts supplied to price changes is zero, by definition, and the response of amounts demanded to price is relatively small, price fluctuations in fresh flowers may be substantial. Inventories cannot supplement below-normal current production; nor can inventories be accumulated to profit from a below-normal current price or from an expected above-normal future price.

At the opposite extreme are the markets for the durable products of agriculture and mining—for example, for wheat or tin. Here storage costs are relatively low, and inventories may easily reach the equivalent of one or even several years' production. Under such circumstances, it is not implausible to argue that rational expectations and inventory adjustments tend to stabilize prices and consumption in the presence of random fluctuations in supply or demand.<sup>3</sup> On the other hand, durable or "permanent" changes in supply or demand alter the equilibrium price, and, almost at once, the actual market price begins to reflect that change.

## II. Price Speculation

However, the existence of large inventories also introduces the possibility of price instability arising from "speculation." Consider, for example, the extreme case of gold, when it is only a commodity, and not used as money. Gold is obviously very durable. And its consumption and its production in any year, or even in any decade, are both very small relative to the size of the total existing stock. The costs of storage are relatively low. It is also nearly indestructible: its services as ornament or store of value are consumed with little or no disappearance or dissipation. Most of the gold that was ever mined presumably still exists, and is potentially available as an addition to the market supply from new production—at prices deemed sufficiently above equilibrium. Gold is currently produced; and its rate of production does

<sup>3</sup>It is therefore somewhat anomalous that, although Muth thus demolished the previous theory of the "cornhog cycle," it is my impression that traces of such price-quantity behavior still persist.

respond to its relative price (and to the expected price that so strongly influences its current price). But a year's production, even at an historically enormous real price, adds very little to the current market supply, compared to that which is available from inventories. Changes in the market price may affect, at least modestly, the quantities of gold demanded: by dentists, by manufacturers of jewelry and objets d'art, or in industry (as a conductor or chemical). But such changes in quantities demanded are likewise small relative to the size of the stock; and even the manufacture of gold objects does not really remove them from the inventory of gold. Indeed, unmined gold (and the shares of gold mining companies) are in effect part of the inventory, too.

Although the price of gold has some effect on its rates of discovery and production, as well as on its consumption (in the sense of its complete and permanent removal from potential market supply), these responses of quantities produced and consumed are truly minimal in comparison with the size of outstanding inventories. Thus, the market price of gold depends mainly on peoples' (or, in the past, on governments') willingness at that price to hold the immense existing inventory of gold. And this means that the current price of gold depends mainly on expectations of its future price. As noted earlier, the current price of a standardized commodity can diverge from any expected future price only by the relatively small costs of storage plus interest (and the cost of assuming the risk of an incorrect expectation). But, in the case of gold, a changed expectation of future price is not soon erased by a changed rate of gold production, nor by an altered level of any consumption that subtracts permanently from the stock. There may always be some normal, long-run equilibrium price of gold that would exactly equate current production and current consumption (in the sense of permanent disappearance). But that normal price does not discipline the actual market price of gold in the way that the normal, long-run equilibrium price of wheat disciplines its current market price. Rather—in the absence of a fixed monetary price of gold—its current price depends overwhelmingly on the current expectation of what its

future price will be. And that future price will depend on the expectation then of its subsequent price.

Such a market is best characterized by borrowing Keynes' parable (pp. 154–160) of the beauty contest, which he used to explain share prices. The price level of shares, he suggested, depends on each market participant's calculation of what the other participants are likely to expect it to be. It is analogous to the contest to choose a beauty queen, he said, in which the prizes go to those who select the candidate for queen thought most beautiful by the largest number of other contestants. Instead of selecting the candidate who is the most beautiful, each contestant tries to calculate which candidate is most likely to be chosen by other contestants. But once each realizes that others will be choosing on the same basis, he is forced to speculate about what the average opinion will be of what the average opinion is. And that speculation can be carried to progressively higher degrees, Keynes suggested. Given no more solid a basis than this for valuing shares, said Keynes, the price of shares can be almost anything—and is likely at times to be highly unstable.

Why does anyone buy gold at \$52 an ounce—or at \$520 an ounce? Why does anyone sell it at that price? Because the buyer expects that the price tomorrow will exceed \$52 by more than the cost of carrying an ounce of gold; and the seller expects that it may be less than \$52 plus carrying costs tomorrow. Each such expectation in turn reflects buyers' and sellers' judgments about what buyers and sellers will be expecting tomorrow for the day after tomorrow. The only other rational reason for buying or selling gold is the pure pleasure of risk assumption—better known as gambling. And there is surely much of that in the gold market. Presumably, pure gambling has little net effect on the price, although it does raise the incomes or employment of brokers.<sup>4</sup>

<sup>4</sup>There is one modest qualification. To the extent that gold prices may move contracyclically vis-à-vis share prices (for example), and are thus held in mixed portfolios with shares in order to reduce risk, the expected return on gold alone might be zero, yet gold might be rationally held (assuming that no other contracyclical asset exists with a positive return.)

Now, what precisely is the basic difference between the wheat and the tin markets, on the one hand, and the gold and the share markets, on the other, that lets us assume in the wheat or tin case that rational expectations will normally *stabilize* prices and consumption in the face either of random or of predictable fluctuations in amounts supplied or demanded, while in the gold case, there is often destabilizing speculation?

The difference obviously has nothing to do with price flexibility and continuous market clearing; for these are trademarks not only of the markets for wheat, tin, and pork bellies, but also for gold (and for General Motors shares). Rather, the difference is that in the wheat and similar cases, the response of *quantities produced and/or quantities consumed* to price changes arising from shifts in supply or demand can be large enough and prompt enough to begin to move the price toward its new equilibrium level within a relevant time period; and that this response therefore comes to be anticipated by market participants. This prompt and substantial stabilizing response of quantities produced and consumed does *not* occur—and therefore it is not *expected* to occur—in the case of gold. As a consequence, in the presence of large inventories, price expectations dominate price-level determination not only in the short term, but even in the moderately long term. The prices so determined are unlikely to contribute either to the *stability* of production or consumption, or to the *equality* of production and consumption. They may thus produce movements in actual inventories, the price effects of which, however, have only modest and long-delayed effects in reversing such movements.

Clearly, wheat and gold are more-or-less extreme examples of a spectrum of durable commodities, capable of storage. At the wheat end of that spectrum, the responses of amounts produced and consumed to price changes are strong and prompt. Market participants come to expect them to occur, and take actions with respect to inventories that tend to stabilize prices around their equilibrium level, including any expected *new* equilibrium level. At the gold end of the spectrum, the responses of amounts produced and consumed to price changes are

weak and slow, while inventories are very large.<sup>5</sup> Participants in such markets thus cannot depend on price movements to be quickly self-limiting. It is therefore not irrational for some or many of them to speculate on the direction of intermediate price movements. Their speculation may often extend and exaggerate price movements that had some origin in changed supply or demand conditions. Or it may extend and exaggerate price movements caused by some random exogenous disturbance, or even by the actions of a few large traders (for example, the Hunts in silver).

Most commodity markets lie between these extremes. In such intermediate markets, price movements may sometimes be—perhaps ordinarily are—in the direction of a price that tends fairly quickly to equate production and consumption. At other times, price movements may develop a cumulative momentum in one direction, which can easily overshoot the current long-run equilibrium price. Occasionally, moreover, speculation can affect even the most stable markets; and speculative fevers can be transmitted from one market to another.

The years 1971 through 1973 provide a clear example of such speculation and of such transmission.<sup>6</sup> Coincidental crop failures in wheat and coarse grains in several of the main producing countries of the world—at a time when world stocks were severely depleted—caused grain prices to soar; this helped to attract public attention to the then popular “Club-of-Rome” fantasies of general resource scarcity, which in turn helped to transmit the speculative fever to still other commodities. In a political environment in which—partly for accidental reasons—macroeconomic policies in a number of major countries were unusually permissive, the speculative rise in commodity prices soon came to be reflected in wage rates and in-

<sup>5</sup>An extreme case, on the side of supply, is that of nonproducible goods (for example, “old master” paintings or old postage stamps) where *all* supply to the market comes from inventory. An earlier version of this paper included a considerable section on prices of such “collectibles.”

<sup>6</sup>For an excellent account of this period, see Barry Bosworth and Robert Lawrence, especially pp. 24–87, and other references given therein.

dustrial costs. Given the inertia of inflation, once it becomes embodied in wage rates and industrial costs; given the fortuitous response of the oil-producing countries when awakened by the international inflation; and given the perhaps appropriate, but surely delayed, responses to inflation by monetary and fiscal policymakers in many of the industrial countries, it is possible to understand why the 1970's turned out to be such a disappointing—even a disastrous—decade. And commodity price speculation played more than a small part in that disaster.

### III. Rational "Price Bubbles"

The fact that rational expectations may not preclude the existence for considerable periods of speculative, self-maintaining price movements toward zero or infinity has recently been recognized in a rapidly expanding theoretical literature of working papers and a few published articles.<sup>7</sup> Indeed, we are seeing a "bubble" of papers on "rational price bubbles." Whether rational expectations are or are not consistent with price bubbles, however, seems to depend entirely on what definition one gives to the "rationality" of rational expectations. One may define as rational *any* expectation that, if generally acted upon, will turn out to be essentially self-confirming. In that case, price bubbles generated by the simplest of extrapolative price expectations may be consistent with rational behavior. On the other hand, one may alternatively define as rational only those expectations that fully incorporate an understanding of "long-run" "market fundamentals": namely, the assumption by each agent of rationally maximizing behavior by each other agent, each responding to observed changes in prices and quantities as would occur in a market model of "long-run equilibrium," although observed through a screen of random, nonserially correlated, short-run disturbances. Actions based on

such expectations thus tend to confirm the expectations; and each confirmation reinforces further reliance up on the long-term equilibrium model. There seems little room for price bubbles in *that* version of rational expectations.

But whether, and to what extent, market fundamentals will, in fact, prevail, surely depends, as I have argued above, on the durability of the commodity, its storage costs, and particularly on the degree to which, and the speed with which, its production and consumption are affected by expected price. Given the world as it is, I suggest that the "fundamentals" do not always and everywhere prevail. Bubbles do occur, and are important.

However, most or all of the new literature on price bubbles appears to deal only with price movements away from, or back towards, the equilibrium price that reflects long-run market fundamentals. It thus implies that, when prices are, for any period, approximately stable, they must therefore be at or close to an equilibrium dictated by market fundamentals—reflecting, at that price, an approximate balance between the expected supply from new production and the expected demand for final consumption. Price bubbles, in contrast, involve an extrapolation of expected price movements (however originating), causing further movement in the same direction.

In my view, however, where the impacts of price on amounts produced and consumed are small, slow, and uncertain, a speculative price—(i.e., a nonequilibrium price) is not always or necessarily a moving price. The price may rest for a considerable period of time at—or fluctuate narrowly around—a level far above (or below) any equilibrium determined by market fundamentals. This situation might represent either the extrapolation of an originally accidental stability, or a standoff between expectations of further rise by some participants and the expectation of fall by others.<sup>8</sup>

<sup>7</sup>See, for example, Olivier Blanchard and Mark Watson; William Brock (1974, 1975); John Cross; Behzad Diba and Herschel Grossman; Robert Flood and Peter Garber; Jo Anna Gray; Maurice Obstfeld and Kenneth Rogoff; and Thomas Sargent and Neil Wallace.

<sup>8</sup>Alternatively, it might reflect the fact that market participants differ in their conceptions of the appropriate model of long-run equilibrium, or receive conflicting information about factors affecting it. Indeed, in

Under circumstances in which uncertainty is overpowering, as Keynes suggested was often the case for stock prices, a price may remain approximately stable for a considerable time on the basis of a tacit "convention," which market participants come to accept. The convention amounts to the assumption that

...the existing state of affairs will continue indefinitely, except in so far as we have specific reasons to expect a change... For, if there exist organized investment markets,... an investor can legitimately encourage himself with the idea that the only risk he runs is that of a genuine change in the news *over the near future*, as to the likelihood of which he can attempt to form his own judgment, and which is unlikely to be very large. For, assuming that the convention holds good, it is only these changes which can affect the value of his investment, and he need not lose his sleep merely because he has not any notion of what his investment will be worth ten years hence.

[*General Theory*, pp. 152-53]

Thus, the price of gold might well, for a considerable period of time, fluctuate narrowly around a level far from the equilibrium price at which amounts produced and consumed would in the long run be equal. And so, although to a considerably lesser extent, could the price of wheat or tin. Yet such a period of approximate stability might at any time be upset by a major, self-maintaining movement in either direction.

#### IV. Macroeconomic Inventory Theory

Let us now turn to some of the *macroeconomic* implications of this microeconomic theory of commodity prices and inventories.

---

a world of constant change, few if any market participants may share the same model of equilibrium—or at least the same and "correct" model. In these situations, most market participants may come to realize that there is no price other than the current one that has any better claim to being considered the equilibrium price.

First, and most important, is to recognize that there is nothing in standard price theory—even when we expand it to take account of speculative demands for inventories, price bubbles, or of the possible transmission of speculative fever from one market to another—that implies that the aggregate stock of inventories, and the rate of aggregate inventory accumulation for an entire economy should exhibit any systematic variation over time. Random disturbances to supply and/or demand for particular commodities or products might lead to periods of net accumulation or net decumulation of inventories of those products. But the aggregation of many time-series, each subject to random, non-serially-correlated variation, produces only a time-series with proportionally smaller variation. Thus, several years of systematic general accumulation of inventories, followed by substantial periods of general decumulation of inventories, finds no basis in conventional price theory.

Yet one of the most obvious macroeconomic facts of life is the existence of pronounced cycles, both in the size of aggregate inventories *and in their rate of accumulation*. Indeed, of all of the conventional subaggregates of real national product, the one that shows the greatest decrease (not merely in percentage but in aggregate dollar amount) between periods of business expansion and of business contraction is, almost invariably, that of the net addition to business inventories. This total typically goes from a large positive amount at business cycle peaks to a large negative amount at business cycle troughs. This difference exceeds the typical cyclical decline in business investment in plant and equipment, in residential construction, in purchases by consumers or governments, or in net exports. The existence of these pronounced cycles in aggregate inventory accumulation has long challenged macroeconomists to develop theories that might explain this phenomenon. The response to this challenge has produced a second and macroeconomic literature about inventories, that is almost completely nonintersecting with microeconomic price theory.

In this macroeconomic inventory literature, prices and price expectations are rare-

ly mentioned. This second literature deals primarily with *work-in-process* and *manufactured product* inventories, including inventories held by wholesalers and retailers—although it may also embrace *commodity* inventories held by manufacturers or processors (but not by producers and dealers). The earliest versions of this literature described inventories as being held, increased, or reduced as the result of production or ordering decisions that were based upon expectations of future changes in *quantities* produced or demanded—expectations generated mainly by recently experienced levels of production or sales. It purported to describe the determinants, for example, of the desired and actual stocks of coal and iron ore held by steel mills; of the desired and actual inventories of finished steel held by mills, steel warehouses, or steel-using manufacturers; of the desired and actual inventories of automobiles held by producers and dealers; and of the desired stocks and actual stocks of steel bedsprings held by department stores. In early versions of this macroeconomic literature, inventory decisions were shown to be capable of having highly destabilizing macroeconomic effects—particularly when expectations of future sales were modelled as responsive in particular ways to experienced levels and changes of sales, while production responded to orders and expected sales.<sup>9</sup>

This literature, of course, belongs to a universe of discourse in which prices (including wage rates) do not clear markets continuously, but rather adjust slowly to evidence of gaps between amounts supplied and demanded. Or they may be “administered” on the basis of working rules of thumb based mainly on costs or on other prices. In this universe of discourse, such pricing rules are not necessarily irrational, given the limited extent of knowledge, and the experience of past instability. And, since prices so established do not typically cause production and

consumption to move quickly into approximate equality, market participants do not expect that prices will so move, and therefore they do not so move. Rather, production and sales are brought toward equality primarily by changes—delayed changes—in planned rates of production or purchase, that reflect observed or expected changes in *quantities* consumed or used in production, and that are designed to restore or to maintain efficient levels of inventories. Indeed, *microeconomic* theories of rational *quantity* adjustments to unplanned differences between production and sales began to be developed in the early 1950’s, and have increasingly supplied a formal basis for macroeconomic inventory theory.<sup>10</sup> These fluctuations in quantities produced or purchased may lead, in the face of relatively inflexible prices and wages, to further self-reinforcing changes in production, purchases, and employment—self-reinforcing through income effects on consumption and “accelerator” effects on investment: perhaps dampened by interest rate effects on investment, to the extent that the supply of money does not accommodate changes in the demand for it.

Indeed, if such patterns of *quantity* responses at the level of the firm to unplanned changes in inventories become standardized—because such patterns do, in time, reverse imbalances between production and sales—it may become entirely rational to expect such patterns of *quantity* movements to occur at a macroeconomic level. Rational expectations based on such “theories” could then lead to inventory, production, and purchase decisions that help to *perpetuate the macroeconomic instability of production, consumption, and inventories*.

Persistent and massive macroeconomic inventory cycles deny either the existence of rational expectations, or, more plausibly, the ability of such expectations to stabilize output and employment. As the persistence of such cycles comes to be expected, and their amplitude to be accepted, it is not surprising that full flexibility of prices comes to be seen

<sup>9</sup>As the name John Muth stands out in the literature on rational expectations, the name of Lloyd Metzler does in the macroeconomic literature on inventory instability. However, more recent work no longer builds directly on Metzler’s. See, for example, Michael Lovell, Martin Feldstein and Alan Auerbach, Alan Blinder (1981, 1982), Blinder and Stanley Fischer.

<sup>10</sup>See Kenneth Arrow, Theodore Harris, and Jacob Marschak; Feldstein and Auerbach; Blinder (1981, 1982).

by many firms as a useless and unprofitable course of conduct.

### V. The Demand for Commodities and the Demand for Labor

I referred earlier to the case (exemplified by fresh flowers) in which the holding of inventories is by definition impossible, and in which the *entire* response to unexpected shifts in supply or demand must be a price change—at least, given perfectly competitive markets. This case may be only a trivial curiosity when we deal with commodities. But it is a matter of considerable importance when we recognize that an inability to store output is, by definition, the case for services—whether they are final products, or are services used in production. The largest class of services, of course, is the labor services used for the production of goods. If the price of labor services was a flexible, continually market-determined price, the price of labor services should respond rather sharply to random shocks to the demand for goods produced using labor. Indeed, prices of labor should, in general, be more variable than prices of commodities, on the reasonable assumption that the supply of labor is less price elastic than the supply of cooperating factors of production.<sup>11</sup>

In fact, of course, the evidence is undeniable that wage rates normally fluctuate far less than do prices. This anomaly seems to require an “institutional” explanation, or, more fundamentally, a theory that explains the institutions that produce these results. Such explanations often run in terms of “set-up,” “transactions,” and “information” costs, that make implicit or explicit contracts mutually beneficial both to workers and to employers, despite the fact that such contracts necessarily imply substantial intermittent unemployment, as well as occasional

labor shortages.<sup>12</sup> An extensive literature has developed along these lines, greatly advanced in Arthur Okun's posthumous *Prices and Quantities: A Macroeconomic Analysis*. The conclusion of Okun's analysis, of course, is that models of the labor market that assume continuous market clearing—and therefore continuous full employment—are erroneous. And, as a result, critiques of government stabilization policy—as powerless to improve economic welfare even when properly used (or unable to damage the national interest when badly used)—necessarily fail.

### VI. A Price-Theoretic Version of the Business Cycle

One well-known critique of stabilization policy is that associated with Robert Lucas and Robert Barro. However, Lucas and Barro accept the reality of a business cycle, involving serially-correlated changes in real output, although they still deny the ability of monetary or fiscal policy to affect the economy.

The price-theoretic approach of Lucas and Barro stands, however, at an opposite pole from that of Okun. For instead of trying to model the imperfect flexibility of wage rates in a real world, wages are sometimes assumed not to exist. In one version (1977) of Lucas' theory of the business cycle, we have a world reminiscent of J. B. Say, in which there are no firms, no hired labor, but only “yeoman farmers.”<sup>13</sup> Thus, there are only prices and profits—no wage rates. Actually, this model could also accommodate simple manufacturers—craftsmen without hired workers—as did the model world of Say. Needless to say, competition is perfect, and information free and nearly complete.

<sup>11</sup>To be sure, the presumed rational expectations response to fluctuations in the demand for commodities, through planned additions to inventories or the drawing-down of inventories, would tend also to stabilize the demand for labor. However, given the costs of commodity storage, it can easily be shown that this response is consistent with the presumption stated in the text: that labor prices should fluctuate more than goods prices.

<sup>12</sup>On the other hand, the explicitly or implicitly contractual nature of the employment relationship not only contributes to the explanation of wage rate insensitivity, but also allows producers greater freedom to use temporary layoffs, along with inventory changes, as a means of adjustment to temporary fluctuations in the demand for output in a manner which may be economically efficient both for workers and for employers. The relationship between inventory behavior and the demand for labor has recently been interestingly explored by Robert Topel.

<sup>13</sup>As Blinder and Fischer have recently characterized the producers in this model.

However, this very simple world has a central bank, or its ruler has a printing press. When bank loans or printing press money are used to finance a public purpose (such as an increment in the Prince's consumption) goods prices are bid up. However, each yeoman farmer or craftsman is likely to confuse the rise in his own selling price as a rise in his price *relative* to prices for the products and services of other yeomen. Such relative price changes occur for various reasons, and are familiar to each of them. Each therefore works harder in the presence of inflation, and produces more, so as to permit him to consume more of the output of his fellow farmers or craftsmen, whose prices he presumes not to have risen. Of course, the Prince has already consumed more of *their* output, too; and their prices, too, have risen. But until each yeoman discovers that the rise in his price is general rather than relative, real GNP increases. Soon, however, each learns that there has been no increase in his real price, and finds only the sour taste of inflation. The boom ends. It can be repeated as soon as the short memories of all citizens are erased.<sup>14</sup>

What can this have to do with the business cycle in a world in which yeoman farmers and craftsmen are replaced by *firms* with hired employees? Very little, I think. For, in even the simplest possible version of this more familiar world, there *have to be two kinds* of prices, not one: prices for goods, and prices for labor. Increased production now requires increased inputs not only from proprietors, but also from their workers. Under standard price theory, in order for employers to wish to hire more labor services, they must believe that product prices have risen relative to wages; but in order for exist-

ing workers to respond to inflation by working longer hours, and other potential workers to enter the labor force, requires that they interpret what happens as a rise in the ratio of wages to prices. It was a nice enough trick to be able to fool each yeoman about his own real price; but it is an even nicer trick for an inflation to fool both employers and workers—in *opposite directions*—about the movement of the real wage paid by one and received by the other!<sup>15</sup>

I suggest that more may be learned about cycles from Knut Wicksell's discussion of business fluctuations, seventy-five years earlier. Wicksell understood that, even in a similarly simple world, there needed to be not merely one kind of price level—nor merely two kinds—but actually three: price levels for goods, for labor, and for loans. And whether or not there was a Prince with a printing press, there were competitive banks, aggressively lending at a flexible price to businesses. Given a plausible (although incompletely specified) lag structure, Wicksell showed that banks' competition in money creation might initially reduce the price for loans; and their continuing money creation might keep it depressed for a time. Through this means, investment, financed by bank credit, could exceed *ex ante* saving, crowding out consumption. The result, of course, was also inflation, unanticipated by all. Once the limits of money creation were reached, the boom collapsed.

At roughly the same time, in another part of Europe, J. A. Schumpeter was describing a rather different source of instability. His bourgeois prince—the innovating entrepreneur—driven by new technological, managerial, or marketing ideas that promised abnormal profits, repeatedly upset the general equilibrium: accommodated, of course, by an elastic banking system. To be sure, his entrepreneurial activities—his “creative destruction”—would soon generate an inflation that had to subside before the next wave of innovation might again disturb the economy.

<sup>14</sup>In another version (Lucas, 1975), there are genuine firms selling products and purchasing labor and capital inputs; but production is scattered among noncommunicating islands, using capital that is immobile and labor that is randomly mobile (between periods), while increments of money are distributed stochastically among the separated markets from period to period. This schema permits changes in money to create fluctuations of real output (of capital goods); but it is not at all clear (to me) why or how the postulated *source* of such “cycles” bears any relationship to the cycles of the “real world.”

<sup>15</sup>The assumption of this mutually inconsistent self-deception appears (at least implicitly) in a number of verbal accounts of a monetarist description of inflation.

Of course, the disturbances of macroeconomic equilibrium that were visualized by Wicksell and Schumpeter, and by Friedrich von Hayek, R. G. Hawtrey, D. H. Robertson, the early Keynes (and so many others) were, in the end, mainly disturbances of the price level, although they normally left a permanent legacy of a larger capital stock, newer technology, and increased human capital.<sup>16</sup> However, most of these pioneers recognized that the price level did not automatically, instantaneously, and proportionately reflect every change in the stock of money or the output of goods; for instance, many recognized that prices did not fall sufficiently promptly in recessions to avoid non-frictional unemployment of labor and/or machines. And such unemployed resources then permitted a subsequent expansion of real output, once the next boom began.

#### VII. Price-Theoretic Models of Aggregate Investment

Most business cycle theories—or, more broadly, most macroeconomic theories of the “medium run”—from the time of Wicksell or even earlier, have thus been built on more-or-less elaborate price-theoretical considerations, related essentially to *investment*; and mainly to fixed investment. For, clearly, it is investment—not consumption or proprietors’ labor—that is the primary source of medium-run macroeconomic disturbance and instability. Although (as noted earlier) sharp fluctuations in the rate of investment in inventories contribute more to U.S. business cycle *recessions* and to the early stages of business recoveries than do fluctuations in plant-and-equipment investment, the latter typically contribute significantly more to *expansions* of GNP from cyclical troughs to peaks than does inventory accumulation. Moreover, in the medium and longer run, fluctuations in inventories are governed by most of the same factors that govern plant and equipment expenditures.

<sup>16</sup>For effective summaries of these and related theories, see Gottfried Haberler. Hayek (and others of the “Austrians”) denied that there were any positive legacies from the boom.

There are many price levels—relative price levels—important for investment in plant, equipment, and normal inventories. They include:

1) The price level for loans: some particular rate of interest, or some average of rates, at which investment can be financed;

2) The *demand*-price level of new capital goods, reflecting their physical productivity, the prices of the goods produced by their use, and the supply-prices of cooperating factors of production;

3) The *supply*-price level for new capital goods, sometimes taken as a function of their rate of production;

4) The *supply*-price and the *demand*-price levels of entrepreneurship and innovation;

5) The *supply*-price levels of risk and/or uncertainty bearing, by investors and lenders;

6) The level and nature of the *prices imposed by government*; taxes and their structure; and

7) The price level of ownership of *existing enterprises*; that is, the level of share prices.

The question is not which among these price levels are *relevant* for investment: all of them (and others) are doubtless relevant to some degree. Rather, the question is: which are “strategically” important? Which are the price levels, the changes in one or more of which relative to another or others are primarily responsible—in the *real* world—for macroeconomic stability or instability, real and nominal?

I can illustrate the wide *variety* of price theoretic explanations of investment by reference to three familiar but quite different examples.<sup>17</sup>

<sup>17</sup>However, not all macroeconomic investment theories have been or are price theoretical. Examples of quantity-theoretic investment theories are the following: (a) Robert Eisner’s adaptation of the accelerator theory, which, in practice, makes real investment, and thus employment and income, depend on fluctuations in current levels of real “permanent” business sales (analogous to permanent income); (b) The works of W. H. L. Anderson, James Duesenberry, and others, whose emphases have mainly been on factors influencing the gross flow of internal funds for investment financing, including tax rates seen as a determinant of internal funds-availability rather than of profitability; and (c) Keynes and George Katona, who stressed political, sociological,

I remind you of Martin Feldstein's complex analyses of the interaction of tax laws and inflation as affecting the profitability of investment. Alternatively, there is Dale Jorgenson's "neoclassical" investment theory, which carefully models the relationship among all of the prices and quantities that enter into dynamic profit-maximizing investment decisions.<sup>18</sup>

As a third example, there is James Tobin's theory, suggested by Keynes, that focuses on the relationship between the aggregate market value of the marketable debt and equity claims against existing enterprises, and the aggregate cost of reproducing, at today's prices, the assets of those enterprises. The ratio of these two global magnitudes has come to be designated as " $q$ ."

I propose to conclude these comments on price-theoretical explanations of investment with a few remarks on the  $q$  theory. The Keynes-Tobin approach makes the volume of aggregate investment dependent upon—indeed, ordinarily taken as proportional to—the excess of  $q$  over 1: where  $q$  is the ratio of the market value of enterprises to the replacement cost of their assets. The market value of enterprises includes the value both of equities and debts, as currently priced in security markets; the replacement cost of their assets is the reproduction cost of existing plant, equipment, and inventories at current price levels.<sup>19</sup>

However formulated in detail, the  $q$  theory of course employs much of the same information that is used in other macroeconomic theories of investment. For example, most

investment theories incorporate a bond-market interest yield, presumably to represent the cost of borrowed long-term funds, or the alternate return on owned funds. Some investment theories may use short-term interest rates, as well, to represent the cost of (or the alternative return on) funds borrowed to finance inventories or other working capital, or to finance the purchase of capital goods while awaiting more favorable rates in the long-term market. The  $q$  theory incorporates interest rates as they affect the value of marketable debt claims against enterprises. This use of interest rates in the  $q$  theory is thus conventional in investment theory.

What is essentially unique about the  $q$  formulation is its inclusion of the market value of stocks: the price level of ownership of corporations. This can be thought of as an indirect measure of the cost (or opportunity cost) of equity funds.<sup>20</sup> Given the typical volatility of share prices, medium-run movements of  $q$  are often or perhaps ordinarily dominated by changes in the prices and thus the market value of outstanding shares.

With that in mind, let me recall to you my earlier discussion of the speculative demand for and supply of storable commodities—gold was the specific example. You recall my borrowing Keynes' parable of the beauty contest to describe such markets. You also recall that this parable was, in fact, Keynes' analogy for the price determination of equities.

In recent years, economists have been paying greatly increased attention to stock prices—a subject largely avoided by economists after the debacle, in and after 1929, both of share prices and of economists' forecasts of share prices. The revival of economists' interest in stocks has involved the development and application of the "capital asset pricing model" and associated "portfolio theories," and, most recently, the attempted application of rational expectations theory to stock prices.

and attitudinal factors influencing the optimism or pessimism of investors or consumers, and thus their spending at given prices, incomes, and interest rates.

<sup>18</sup>To be sure, in Jorgenson's own applications of his theory, he concentrates on levels of and changes in tax-law depreciation provisions, tax credits, and the tax rates that affect after-tax profitability: even to the extent of taking—in empirical work—the relevant interest rate, used to discount future cost and income streams, as a constant over time!

<sup>19</sup>The linearity of the relationship between investment and  $q$  is, of course, arbitrary, as is the use of the ratio less one rather than the algebraic difference between market value and replacement cost. Several variants are thus possible.

<sup>20</sup>Stock prices can also be thought of merely as an index of sentiment, optimism, or "animal spirits," which seem so significantly to affect both the stock market and real investment. Although Keynes stressed *this* influence of stock prices, that is not their role in Tobin's analysis.

However, Robert Shiller's recent paper "Do Stock Prices Move Much Too Much to be Justified by Subsequent Changes in Dividends?" appears to demolish the possibility that movements of U.S. stock prices can be explained by the rational expectations of share holders.<sup>21</sup> For over a century, real stock prices appear to have fluctuated far more than any plausible change in the rational expectation of real dividends or earnings. Shiller does not attempt to characterize the source of the greatly excessive volatility of stock prices. But, surely, it is possible that speculative price bubbles, upward or downward, based upon the extrapolation of nominal share-price levels and movements, and on the effort to profit (or to avoid loss) from such movements, supply some part of the explanation. Another piece of the explanation for recent stock prices may well be Franco Modigliani's and Richard Cohn's "inflation-illusion" theory—itself clearly inconsistent with rational expectations.

The rational expectations theory, at least as applied to individual commodities, assumes that there exists, at all times, some long-run equilibrium price—that price which would equate amounts produced and consumed. Market participants can and do have at least a rough idea of what this price is; moreover, their individual estimates of this price are basically similar, for all have essentially the same information, and use essentially the same model of how that price is determined.

Those who attempt to apply the concept of rational expectations to the stock market rarely state their understanding of what is the *equilibrium* price level for shares—or of a particular share—that it would be rational for participants to expect. Is it the level of share prices at which net new issues of shares would be zero? Or would it be the level of share prices at which net private investment would be zero? The level at which net private investment would equal net saving at full employment? The level at which investment might maintain the economy on a Golden

Rule growth path? Or is there some concept of a less-fundamental, less-long-run equilibrium level of share prices which market participants can be expected to understand and quantitatively approximate? If there is no such concept of equilibrium share-price level, how can there be a unique rational expectation of share prices?

A simpler course is to admit that we have no very precise concept of an equilibrium level of share prices, but to argue that we can nevertheless predict the direction and rate of movement over time of that equilibrium, whatever it may be. The rate of movement might, for example, be expected to approximate the rate of growth of profits or of dividends per share. (This was Shiller's device.) This may suffice for rough tests of the *ex post* rationality of historical share price movements. But it offers essentially zero guidance to the purchaser or seller of a particular stock at a particular time. He must guess whether the prices of particular stocks will rise or fall, over some less than infinite horizon, from where they stand today. Is it strange that he is more concerned with the correctness of his guess about what other buyers and sellers are expecting will happen to prices of those particular shares, and to the market averages? Is there a better description than the "beauty contest" parable?

Actually, Keynes' description of the determination of stock prices is far more detailed and substantive than I have here described it, and I recommend its occasional rereading. However, there is no reason to suppose that Keynes in 1935 had the last word on stock prices. Henry Wallich, an acute observer, who began his distinguished career on Wall Street, has recently written about the "Radical Revisions of the Distant Future" that occur from time to time in the stock market, and for which one cannot find rational explanation, in any narrow sense of that term. His explanation for at least a part of the most recent "radical revision"—since about 1973—runs in terms that I can perhaps best describe as "sociological":

One might guess that [the reasons for this massive change in investment attitudes] have something to do with

<sup>21</sup>For an important related paper, reaching similar conclusions by different techniques, see William Brainard, John Shoven, and Laurence Weiss.

the professionalization of the securities business. Very likely this tends to homogenize views, increasing the herd instinct among bulls and bears, respectively. The rewards/penalties system for professionals works in that direction. It is dangerous to be wrong in support of an unpopular cause....

Professionals have made the market efficient, in the narrow sense that there is nothing of predictive value to be learned from the past data of the market. It is far from clear that they have made it rational. There may be something to be learned from the history of mass delusions in the market after all. [p. 38]

Modigliani and Cohn's suggestion may be one such systematic mass delusion.

As one close to retirement, and one of quite a number of us (I suspect) who have left most of our retirement resources tied up in CREF, I would of course, be pleased if stock prices should return closer to their average past relationship to earnings or dividends. But I would not, I confess, find it a confirmation of the rationality of my own expectations. Personally, I long ago decided that being an economist was not merely no advantage, but probably a disadvantage in the security markets; and I have never personally participated in them. I have been fearful, I suppose, of succumbing, myself, to the herd instincts that I seem to observe there. The broker who tells his customers that Joe Granville is stupid, but that they must pay attention to him because others will, both assumes and encourages behavior only one level removed from that of lemmings.

Because stock prices are not fully rational, either in the large or even in the small, sharp-eyed members of several generations of my graduate students learned (not from me) to support themselves in reasonable comfort by playing on trivial systematic anomalies that they had found in share price movements. They succeeded, presumably, by acting exactly as others do, but a trifle sooner.

Indeed, the lemming instinct affects participants in other markets. One current example is that of the international bankers,

whose loans to particular *LDCs* were safe because—but only so long as—other international bankers recognized exactly the same source of safety.

Whether one describes investment decisions in terms of the  $q$  formulation or in some other way, prices in security markets necessarily affect the volume of aggregate investment. And, because such prices are clearly not fully rational, investment is a potential and actual source of exogenous disturbance of macroeconomic equilibrium; and successful government stabilization policies are not, by definition, precluded.

Our stroll along the border between micro- and macroeconomics comes to an end. What can we conclude? I conclude that the rational expectations model of economic behavior adequately describes an important range of economic activities, where prices adjust smoothly and efficiently to clear markets and to stabilize production and consumption over time. But rational expectations do not adequately explain other kinds of markets, where *speculative* prices may systematically tend to overshoot changing equilibrium levels: nor yet another kind—including most labor markets and many “customers markets”—where price changes tend systematically to undershoot changing equilibrium levels, whether because of an inability to develop efficient long-term contracts, the existence of bilateral monopoly, or merely because of the rapid pace of unpredictable technological, institutional, or other exogenous change. All of the resulting aberrant forms of *microeconomic* behavior may in some sense be individually “rational”; yet their *macroeconomic* effects are often perverse.

Nevertheless, we economists do our best to understand our world, and to discover those dependable regularities of behavior—whether or not fully rational—that provide the basis for economics *theories*, which we can then use to prescribe *policies*, whether these policies are of *laissez-faire* or of selective intervention. All forms of dependably regular behavior that we seek to discover and to describe—and not merely those that are fully rational—are equally important parts of the *social science* of Economics, that Marshall once defined simply as “a study of

mankind in the ordinary business of life; ...that part of individual and social action which is most closely connected with the attainment and use of the material requisites of wellbeing" (p. 1, emphasis added).

## REFERENCES

- Ackley, Gardner, *Macroeconomic Theory*, New York: Macmillan, 1961.
- Anderson, W. H. L., "Business Fixed Investment: A Marriage of Fact and Fancy," in Robert Ferber, ed., *The Determinants of Investment Behavior*, New York: Columbia University Press, 1967, 413-25.
- Arrow, Kenneth, Harris, Theodore and Marschak, Jacob, "Optimal Inventory Policy," *Econometrica*, July 1951, 19, 250-72.
- Barro, Robert J., "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, 67, 101-15.
- , "Unanticipated Money, Output, and the Price Level in the United States," *Journal of Political Economy*, August 1978, 86, 549-80.
- Blanchard, Olivier J. and Watson, Mark W., "Bubbles, Rational Expectations and Financial Markets," Discussion Paper Number 877, Harvard Institute of Economic Research, January 1982. (Also circulated as National Bureau of Economic Research Working Paper No. 945, July 1982.)
- Blinder, Alan S., "Retail Inventory Behavior and Business Fluctuations," *Brookings Papers on Economic Activity*, 2:1981, 443-505.
- , "Inventories and Sticky Prices: More on the Microfoundations of Macroeconomics," *American Economic Review*, June 1982, 72, 334-48.
- and Fischer, Stanley, "Inventories, Rational Expectations, and the Business Cycle," *Journal of Monetary Economics*, November 1981, 8, 277-304.
- Bosworth, Barry P. and Lawrence, Robert Z., *Commodity Prices and the New Inflation*, Washington: The Brookings Institution, 1982.
- Brainard, William C., Shoven, John B. and Weiss, Laurence, "The Financial Valuation of the Return to Capital," *Brookings Papers on Economic Activity*, 2:1980, 453-502.
- Brock, William A., "Money and Growth: The Case of Long-Run Perfect Foresight," *International Economic Review*, October 1974, 15, 750-77.
- , "A Simple, Perfect Foresight Monetary Model," *Journal of Monetary Economics*, April 1975, 1, 133-50.
- Calvo, Guillermo, "On Models of Money and Perfect Foresight," *International Economic Review*, February 1979, 20, 83-103.
- Cross, John G., "On Baubles and Bubbles," undated, pp. 1-22.
- Diba, Behzad T. and Grossman, Hershel, I., "Rational Asset Price Bubbles," Economics Working Paper No. 81-83, Brown University, March 1982.
- Duesenberry, James S., *Business Cycles and Economic Growth*, New York: McGraw-Hill, 1958.
- Eisner, Robert, "A Permanent Income Theory of Investment: Some Empirical Explorations," *American Economic Review*, June 1967, 57, 363-90.
- Feldstein, Martin, S., "Inflation and the Stock Market," *American Economic Review*, December 1980, 70, 839-47.
- and Auerbach, Alan, "Inventory Behavior in Durable Manufacturing: The Target-Adjustment Model," *Brookings Papers on Economic Activity*, 2:1976, 351-96.
- Flood, Robert P. and Garber, Peter M., "Market Fundamentals versus Price Level Bubbles: The First Tests," *Journal of Political Economy*, August 1980, 88, 745-70.
- Gray, Jo Anna, "Dynamic Instability in Rational Expectations Models: An Attempt to Clarify," International Finance Discussion Papers, No. 197, January 1982.
- Haberler, Gottfried, *Prosperity and Depression*, 3rd ed., Geneva: Economic Intelligence Service, League of Nations, 1941.
- Irvine, F. Owen, "Retail Inventory Investment and the Cost of Capital," *American Economic Review*, September 1981, 71, 633-48.
- Jorgenson, Dale W., "Capital Theory and Investment Behavior," *American Economic Review Proceedings*, May 1963, 53, 247-68.
- , "The Theory of Investment Behavior," in Robert Ferber, ed., *The Determinants of Investment Behavior*, New York: Columbia University Press, 1967, 1-22.

- nants of Investment Behavior, New York: Columbia University Press, 1967, 1129-56.
- Katona, George, *Psychological Economics*, Amsterdam: Elsevier, 1975.
- Keynes, John Maynard, *The General Theory of Employment Interest and Money*, New York: Harcourt Brace and Co., 1936.
- Lovell, Michael C., "Manufacturers' Inventories, Sales Expectations, and the Acceleration Principle," *Econometrica*, July 1961, 29, 293-314.
- Lucas, Robert E., Jr., "Understanding Business Cycles," in K. Brunner and A. Meltzer, eds., *Carnegie-Rochester Series on Public Policy*, Vol. 5, Amsterdam: North-Holland, 1977, 7-29.
- , "An Equilibrium Model of the Business Cycle," *Journal of Political Economy*, December 1975, 83, 1113-144.
- Marshall, Alfred, *Principles of Economics*, 8th ed., London: Macmillan, 1920.
- Metzler, Lloyd A., "The Nature and Significance of Inventory Cycles," *Review of Economics and Statistics*, February 1947, 29, 1-5.
- Modigliani, Franco and Cohn, Richard A., "Inflation, Rational Valuation, and the Market," *Financial Analysts Journal*, March-April 1979, 35, 24-44.
- Muth, John, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, 29, 315-35.
- Obstfeld, Maurice and Rogoff, Kenneth, "Speculative Hyperinflations in Maximizing Models: Can We Rule Them Out?," Working Paper No. 855, National Bureau of Economic Research, February 1982, 1-27.
- Okun, Arthur, *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.
- Sargent, Thomas J. and Wallace, Neil, "The Stability of Models of Money and Growth with Perfect Foresight," *Econometrica*, November 1973, 41, 1043-48.
- Schumpeter, Joseph A., *The Theory of Economic Development*, R. Opie, transl., Cambridge: Harvard University Press, 1939. (Original publication in German, 1912.)
- Shiller, Robert J., "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, 71, 421-36.
- Tobin, James and Brainard, William C., "Asset Markets and the Cost of Capital," in Bela Balassa and Richard Nelson, eds., *Economic Progress, Private Values and Public Policy, Essays in Honor of William Fellner*, Amsterdam: North-Holland, 1977.
- Topel, Robert H., "Inventories, Layoffs, and the Short-Run Demand for Labor," *American Economic Review*, September 1982, 72, 769-87.
- Wallich, Henry C., "Radical Revisions of the Distant Future," *Journal of Portfolio Management*, Fall 1979, 1, 36-38.
- Wicksell, Knut, *Lectures on Political Economy*, L. Robbins, transl., London: Routledge and Kegan Paul, 1934. (Original publication in Swedish, 1901-06.)

# Why Do Companies Pay Dividends?

By MARTIN FELDSTEIN AND JERRY GREEN\*

The nearly universal policy of paying substantial dividends is the primary puzzle in the economics of corporate finance. Until 1982, dividends were taxed at rates varying up to 70 percent and averaging nearly 40 percent for individual shareholders. In contrast, retained earnings imply no concurrent tax liability; the rise in the share value that results from retained earnings is taxed only when the stock is sold and then at least 60 percent of the gain is untaxed.<sup>1</sup> In spite of this significant tax penalty, U.S. corporations continue to distribute a major fraction of their earnings as dividends; during the past fifteen years, dividends have averaged 45 percent of real after-tax profits. In effect, corporations voluntarily impose a tax liability on their shareholders that is currently more than \$10 billion a year.<sup>2</sup>

Why do corporations not eliminate (or sharply reduce) their dividends and increase their retained earnings?<sup>3</sup> It is, of course,

arguable that if all firms were to adopt such a policy, it would raise the aggregate level of investment and therefore depress the rate of return on capital.<sup>4</sup> But any individual firm could now increase its retained earnings without having to take less than the average market return on its capital if it used the additional funds to diversify into new activities or even to acquire new firms.

Several different possible resolutions of the dividend puzzle have been suggested. In reality there is probably some truth to all of these ideas, but we believe that, even collectively, they have failed to provide a satisfactory explanation of the prevailing ratio of dividends to retained earnings. It is useful to distinguish five kinds of explanations.

First, there is the desire on the part of small investors, fiduciaries, and nonprofit organizations for a steady stream of dividends with which to finance consumption. Although the same consumption stream might be financed on a more favorably taxed basis by periodically selling shares, it is argued that small investors might have substantial transaction costs and that some fiduciaries and nonprofit organizations are required to spend only "income" and not

\*Harvard University and the National Bureau of Economic Research. This paper is part of the NBER study of Capital Formation and its research program on Business Taxation and Finance. We have benefited from discussion of this work with Alan Auerbach, David Bradford, John Flemming, Mervyn King, Lawrence Summers, and other participants in the NBER's 1979 summer institute. We are grateful for financial support to the NBER and the National Science Foundation. The views expressed here are our own and are not those of the NBER or Harvard University.

<sup>1</sup>Current law allows 60 percent of the gain to be excluded. This has the effect of taxing realized capital gains at only 40 percent of the regular income tax rate. When shares that are obtained as a bequest are sold, the resulting taxable income is limited to 40 percent of the rise in the value of the shares since the death of the previous owner.

<sup>2</sup>There would of course be no problem in explaining the existence of dividends if there were no taxes. The analysis of Franco Modigliani and Merton Miller (1958) shows that without taxes, dividend policy is essentially irrelevant since shareholders can in principle offset any change in dividend policy by buying or selling shares. Even in the Modigliani-Miller world, the stability of dividend rates would require explanation.

<sup>3</sup>There is also in principle the possibility of repurchasing shares instead of paying dividends. The pro-

ceeds received by shareholders would be taxed at no more than the capital gains rate and therefore at no more than 40 percent of the rate that would be paid if the same funds were distributed as dividends. There are however significant legal impediments to a systematic repurchase policy. Regular periodic repurchases of shares would be construed as equivalent to dividends for tax purposes. Sporadic repurchases would presumably avoid this, but would subject managers and directors to the risk of shareholder suits on the grounds that they benefited from insider knowledge in deciding when the company should repurchase shares and whether they as individuals should sell at that time. British law forbids the repurchase of shares. The present paper assumes that frequent repurchases would be regarded as income and therefore focuses on the choice between dividends and retained earnings. The possibility of postponed and infrequent share repurchases is expressly considered.

<sup>4</sup>The greater retained earnings could also partly or wholly replace debt finance.

"principal." However, transaction costs could be reduced significantly if investors sold shares less frequently. Fiduciaries and nonprofit organizations can often eliminate any required distinction between income and principal.

Merton Miller and Myron Scholes (1978) have offered the ingenious explanation that the current limit on interest deductions implies that there is no marginal tax on dividends. Under current tax law, an individual's deduction for investment interest (i.e., interest other than mortgage and business interest) is limited to investment income plus \$10,000. An extra dollar of dividend income raises the allowable interest deduction by one dollar. For a taxpayer for whom this constraint is binding, the extra dollar of dividends is just offset by the extra dollar of interest deduction, leaving taxable income unchanged. Although Miller and Scholes discuss how the use of tax-exempt annuities "should" make this constraint binding for all individual investors, in reality fewer than one-tenth of 1 percent of taxpayers with dividends actually had large enough interest deductions to make this constraint binding.<sup>5</sup> Moreover, since the limit on interest deductions was only introduced in 1969, the Miller-Scholes thesis is irrelevant for earlier years.

A more plausible explanation is that dividends are required because of the separation of ownership and management. According to one form of this argument, dividends are a signal of the sustainable income of the corporation: management selects a dividend policy to communicate the level and growth of real income because conventional accounting reports are inadequate guides to current income and future prospects.<sup>6</sup> While this theory remains to be fully elaborated, it does suggest that the steadiness (or safety) of the dividend, as well as its average level, might

be used in a dynamic setting. The dividend tax of more than \$10 billion does seem to be an inordinately high price to pay for communicating this information; a lower payment ratio might convey nearly the same information without such a tax penalty. Closely related to the signalling idea is the notion that shareholders distrust the management and fear that retained earnings will be wasted in poor investments, higher management compensation, etc. According to this argument, in the absence of taxation shareholders would clearly prefer "a bird in hand," and this preference is strong enough to pressure management to make dividend payments even when this involves a tax penalty. If investors would prefer dividends to retained earnings because of this distrust, it is hard to understand why there is not pressure for a 100 percent dividend payout.<sup>7</sup>

Alan Auerbach (1979), David Bradford (1979), and Mervyn King (1977) have developed a theory in which positive dividend payments are consistent with shareholder equilibrium because the market value per dollar of retained earnings is less than one dollar. More specifically, if  $\theta$  is the tax rate on dividends and  $c$  is the equivalent accrual tax rate on capital gains,<sup>8</sup> the net value of one dollar of dividends is  $1 - \theta$ , while the net value of one dollar of retained earnings is  $(1 - c)p$  where  $p$  is the rise in the market value of the firm's shares when an extra dollar of earnings is retained, that is,  $p$  is the share price per dollar of equity capital. Auerbach, Bradford, and King point out that shareholders will be indifferent between dividends and retained earnings if the share price per dollar of equity capital is  $p = (1 - \theta)/(1 - c) < 1$ . At any other value of  $p$ , shareholders would prefer either no dividends or no retained earnings but at  $p = (1 - \theta)/(1 - c)$  any value of the dividend

<sup>5</sup>Daniel Feenberg (1981) uses a large sample of actual tax returns to estimate the number of dividend recipients affected by the interest income deduction limitation. He finds that in 1977 only 2.5 percent of dividend income goes to constrained taxpayers.

<sup>6</sup>For a development of this view, see Sudipto Bhattacharya (1979), Roger Gordon and Burton Malkiel (1979), and Stephen Ross (1977).

<sup>7</sup>The argument that dividends reflect the separation of ownership and management appears to be supported by the fact that closely held companies pay little or no dividends. However, such companies can usually achieve a distribution of funds as management salary which is deductible.

<sup>8</sup>The equivalent accrual tax rate on capital gains is the present value of the tax liability that will eventually be paid, per dollar of dividend income.

payout rate would be equally acceptable. Moreover, in the context of their model, the share price will satisfy this value of  $p$  when shares sell at the present value of after-tax dividends. In short, they argue that the existence of dividends is appropriate if the value of retained earnings capitalizes the tax penalty on any eventual distribution.

This line of reasoning is clearly important but raises several problems. First, it has been argued<sup>9</sup> that an equilibrium in which  $p$  is less than one is incompatible with new equity finance by the firm. While it is clearly inconsistent for firms to pay dividends and sell shares at the same time (except if dividends are paid for some of the other reasons noted above), the theory is not incompatible with firms having some periods when  $p \geq 1$  and new equity is sold and other periods when  $p < 1$  and dividends are paid but shares are not sold. In any case, new equity issues by established companies (outside the regulated industries where special considerations are applicable) are relatively rare.

A more important problem with the Auerbach-Bradford-King theory is that it is based on the premise that funds can never be distributed to shareholders in any form other than dividends. This implicitly precludes the possibility of allowing the company to be acquired by another firm or using accumulated retained earnings to repurchase shares. Either of these options permits the earnings to be taxed as capital gains after a delay.<sup>10</sup> The theory that we develop in the present paper explicitly recognizes this possibility.

A further difficulty with the theory is that any payout rate is consistent with equilibrium and therefore gives no reason for the observed stability of the payout rate over time for individual companies and for the aggregate. Although such stability could be explained by combining the Auerbach-

King-Bradford model with some type of signalling explanation, our own analysis based purely on considerations of risk indicates that the payout rate is determinate and that it is likely to be relatively insensitive to fluctuations in annual earnings. (A more explicit dynamic analysis would be necessary to confirm this conclusion.)

The most serious problem with the Auerbach-Bradford-King hypothesis is the implicit assumption that all shareholders have the same tax rates ( $\theta$  and  $c$ ). In reality, there is substantial variation in tax rates and therefore in the value of  $p = (1 - \theta)/(1 - c)$  that is compatible with a partial dividend payout. For individuals in the highest tax bracket,  $\theta = 0.7$  and the dividend-compatible  $p$  is approximately 0.33;<sup>11</sup> for tax-exempt institutions, the corresponding value is one. The Auerbach-Bradford-King concept of shareholder equilibrium implies that, at any market value of  $p$ , almost all shareholders will prefer either no dividends or no retained earnings, depending on whether the market value of  $p$  was greater than or less than their own values of the ratio  $(1 - \theta)/(1 - c)$ . This condition would cause market segmentation and specialization; some firms would pay no dividend while others would have no retained earnings and each investor would own shares in only one type of firm. Such specialization and market segmentation is clearly counterfactual. Our own current analysis emphasizes the diversity of shareholder tax rates and shows that this is a key to understanding the observed policy of substantial and stable dividends.

In our 1979 paper with Eytan Sheshinski, we studied the long-run growth equilibrium of an economy with corporate and personal taxes. In this context, dividends appear as the difference between after-tax profits and the retained earnings that are consistent with steady-state growth and with the optimal debt-equity ratio. This limits aggregate retained earnings and implies positive aggregate dividends, but does not explain why

<sup>9</sup>See, for example, Gordon and Malkiel.

<sup>10</sup>Such infrequent share repurchases are very different from a systematic program of substituting regular repurchases for dividends. They do not risk the adverse tax consequence referred to above and, unlike continuous repurchases in lieu of dividends, involve a different growth of equity.

<sup>11</sup>This is based on tax rates for 1981 and assumes that postponement and the stepped-up basis at death reduce the accrual equivalent capital gains tax to 10 percent.

each firm will choose to pay positive dividends rather than to grow faster than the economy's natural rate. We suggested that each firm is constrained by the fact that more rapid growth would increase its relative size, thereby making it riskier and reducing the market price of its securities. An explicit model of this relation between size and the "risk discount" was not presented in that paper, but is one of the basic ideas of the general equilibrium analysis that we present here. Unlike the previous paper, the present analysis will not look at properties of the long-run steady state, but will examine microeconomic choice in a one-period model.

The idea of shareholder risk aversion as a limit to a firm's growth and the existence of shareholders in diverse tax situations are the two central components of the analysis developed in the present paper. We consider an economy with two kinds of investors: taxable individuals and untaxed institutions (like pension funds and nonprofit organizations).<sup>12</sup> Firms can distribute profits currently as dividends, or retain them, grow larger, and ultimately distribute these funds to shareholders as capital gains.<sup>13</sup> In the absence of uncertainty, these assumptions would lead to segmentation and specialization. The taxable individuals would invest only in firms that pay no dividends even though, *ceteris paribus*, they prefer present dollars to future dollars while untaxed institutions would invest only in firms that retain no profits. In this equilibrium the share price per dollar of retained earnings would in general be less than one. This type of equilibrium with segmentation and specialization is not observed because of uncertainty. Because investors regard each firm's return as both unique and uncertain, they wish to diversify their investment. We show in this paper that each firm can in general maximize its share price by attracting both types of investors, and that

this requires a dividend policy of distributing some fraction of earnings as dividends. Only in the special case of little or no uncertainty or of a limited ability to diversify risks can the equilibrium be of the segmented-market form.

The first section of the paper presents the basic model of dividend behavior in a two-firm economy with two classes of investors. Some comparative statics of the resulting equilibrium are developed in Section II. The third section examines the special case in which the two firms have equal expected yields and equal variances. Despite the diversity of taxpayers, both firms choose the same dividend rate. In Section IV, the symmetry of this equilibrium is contrasted with the segmentation and specialization that can arise with riskless investments, or with risk-neutral individuals. There is a final concluding section that suggests directions for further work.

### I. Dividend Behavior in a Two-Company Economy

Our analysis of corporate dividend behavior uses a simple one-period model. At the beginning of the period, each firm has one dollar of net profits that must be divided between dividends and earnings. The firms announce their dividend policies and trading then takes place in the shares. The firms use the amounts that they have retained to make investments in plant and equipment. At the end of the period, the uncertain returns on these investments are realized and the companies are liquidated. All of the end-of-period payments are regarded as capital gains rather than dividends and will be assumed to be untaxed.

There are two kinds of investors in the economy. Households (denoted by a subscript  $H$ ) are taxed at rate  $\theta$  on dividend income but pay no tax on capital gains. Institutions (denoted by a subscript  $I$ ) pay no taxes on either dividends or capital gains. At the beginning of the period, the two types of investors own the following numbers of shares in both companies:  $\bar{s}_{H1}$ ,  $\bar{s}_{H2}$ ,  $\bar{s}_{I1}$ , and  $\bar{s}_{I2}$ , where subscripts 1 and 2 indicate the companies. For notational simplicity, we normalize the number of shares in each com-

<sup>12</sup>The same reasoning would apply if we consider "low-tax rate" and "high-tax rate" individuals. See Feldstein and Joel Slemrod (1980) for the application of such a classification to analyzing the effect of the corporate tax system.

<sup>13</sup>This future capital gain distribution could be the result of the firm's shares being acquired by another firm or of a share repurchase by the firm itself.

pany at 1. After the companies announce their dividend policies, the investors can sell their shares (at prices determined in the market that depend on the firms' dividend policies) and can buy other shares. Investors can also place some of the proceeds of their share sales in a riskless asset or can spend those funds on consumption; each dollar invested in this riskless asset has an end of period value of  $R$ . We assume, however, that investors may not sell shares short. Both types of investors prefer present dollars to future dollars; one present dollar (obtained either as after-tax dividends or from the sale of shares) is worth  $R$  dollars. Although  $R$  might be expected to differ between households and institutions, we shall assume the same  $R$  for both groups.

Each firm has an initial amount of one dollar available for distribution and retention. Company  $i$  pays dividend  $d_i$  at the beginning of the period and therefore invests amount  $1 - d_i$ . The end-of-period of company  $i$  ( $i = 1, 2$ ) is  $r_i$  per dollar of funds that are retained and invested; the rate of return on the firm's capital is thus  $r_i - 1$ .<sup>14</sup> The expected value of this uncertain return is  $r_i^e$  and its variance is  $\sigma_{ii}$ . The covariance of the returns of the two firms is  $\sigma_{12}$ . In the analysis that follows, we consider the general case in which the yields and variances are unequal. We then examine in detail the character of the equilibrium in the case in which the mean yields and variances of the two firms are identical. We show that in this situation, the degree of uncertainty (as measured by the common variance) and the opportunity for effective diversification (as measured by the correlation between the returns) determine whether both companies pay dividends and are owned by both types of investors or there is market segmentation in which one company pays no dividends and is owned by the household investors.

Our strategy of analysis is as follows. We first derive the share demand equations for the two types of investors. These demands depend on the prices of the shares and on

their stated dividend policies. We then use the fact that the available number of shares of each type of stock is fixed to calculate the price functions. The price of each type of share depends in general on the dividend policy of that firm and of the other type of firm. We assume that firms select the dividend policy that maximizes the firm's value, that is, that maximizes the price per share.<sup>15</sup> This maximization yields the optimal dividend for each firm. When these dividend values have been obtained, we shall examine the characteristics of the equilibrium and the comparative static response to changes in the tax rate.

#### A. Investors' Demands for Shares

We derive each investor's demand functions for shares by maximizing the investor's expected utility subject to the wealth constraint implied by the investor's initial shareholdings and the equilibrium share prices. We assume that the investors' utility functions are quadratic and focus our attention on the role of taxes by assuming that all investors have exactly the same utility function. The nature of the utility function implies that the demand for each type of share is independent of the individual's wealth; we can therefore derive aggregate demand functions for each type of shareholder by treating all of the investors of each type as if they were a single investor.

Consider first the investment problem of the households. If the market equilibrium share prices for the two companies are  $p_1$  and  $p_2$ , the value of their initial portfolio is  $p_1 \bar{s}_{H1} + p_2 \bar{s}_{H2}$ . The initial wealth is exchanged for  $s_{H1}$  shares of company 1,  $s_{H2}$  shares of company 2, and  $z$  dollars of the monetary asset. The new portfolio must satisfy the wealth constraint:

$$(1) \quad p_1 \bar{s}_{H1} + p_2 \bar{s}_{H2} = p_1 s_{H1} + p_2 s_{H2} + z_H.$$

<sup>14</sup>We assume that firms do not borrow and that the stochastic return per dollar of investment does not depend on the amount that is invested.

<sup>15</sup>Maximizing the share price is Pareto efficient, but not uniquely optimal. There are other plausible criteria by which management might in general decide its dividend policy even in a one-period model such as the current one, for example, majority voting of the shareholders.

391  
001

P 7830

With dividend payouts of  $d_1$  and  $d_2$ , the households' total after-tax funds at the beginning of the period are  $(1-\theta)d_1s_{H1} + (1-\theta)d_2s_{H2} + z_H$ . The additional funds received at the end of the period are the uncertain amount  $(1-d_1)s_{H1}r_1 + (1-d_2)s_{H2}r_2$ . Combining these two with each dollar of beginning-of-period funds equivalent to  $R$  dollars of the end-of-period funds yields the argument of the household's utility function:

$$(2) \quad W_H = R(1-\theta)[s_{H1}d_1 + s_{H2}d_2] \\ + Rz_H + s_{H1}(1-d_1)r_1 \\ + s_{H2}(1-d_2)r_2.$$

The quadratic character of the utility function implies that expected utility can be written as a linear combination of the mean and variance of  $W_H$ :

$$(3) \quad E[U(W_H)] = E(W_H) - 0.5\gamma \cdot \text{var}(W_H),$$

where  $\gamma > 0$  is a measure of risk aversion (and the 0.5 is introduced to simplify subsequent calculations). Equation (2) implies that

$$(4) \quad E(W_H) = R(1-\theta)(s_{H1}d_1 + s_{H2}d_2) \\ + Rz_H + s_{H1}(1-d_1)r_1^e \\ + s_{H2}(1-d_2)r_2^e$$

and

$$(5) \quad \text{var}(W_H) = s_{H1}^2(1-d_1)^2\sigma_{11} \\ + s_{H2}^2(1-d_2)^2\sigma_{22} \\ + 2s_{H1}s_{H2}(1-d_1)(1-d_2)\sigma_{12}.$$

The households' optimum portfolio is found by maximizing equation (3) subject to the constraint of equation (1).<sup>16</sup> The first-order conditions for maximizing expected

utility are

$$(6) \quad 0 = R(1-\theta)d_1 + (1-d_1)r_1^e \\ - Rp_1 - \gamma[s_{H1}(1-d_1)^2\sigma_{11} \\ + s_{H2}(1-d_1)(1-d_2)\sigma_{12}]$$

and

$$(7) \quad 0 = R(1-\theta)d_2 + (1-d_2)r_2^e \\ - Rp_2 - \gamma[s_{H2}(1-d_2)^2\sigma_{22} \\ + s_{H1}(1-d_1)(1-d_2)\sigma_{12}].$$

Collecting terms, we may write the households' pair of demand equations as

$$(8) \quad \gamma \begin{bmatrix} (1-d_1)^2\sigma_{11} & (1-d_1)(1-d_2)\sigma_{12} \\ (1-d_1)(1-d_2)\sigma_{12} & (1-d_2)^2\sigma_{22} \end{bmatrix} \\ \cdot \begin{bmatrix} s_{H1} \\ s_{H2} \end{bmatrix} = \begin{bmatrix} R(1-\theta)d_1 + (1-d_1)r_1^e - Rp_1 \\ R(1-\theta)d_2 + (1-d_2)r_2^e - Rp_2 \end{bmatrix}$$

or, in matrix notation,

$$(9) \quad \gamma A s_H = a_H - R p,$$

where the elements of  $A$  and  $a_H$  are clear from (8). If the matrix  $A$  is not singular, (8) can be solved for the share demands  $s_H$ . It is important to note that  $A$  is singular when either stock is riskless or when the correlation between the two yields is one; in either case, holding a mixed portfolio does not achieve any reduction in risk. The optimal portfolio in this case is an investment in only one type of stock. More generally, when the variances are small or the correlation high, the solution of equation (9) may imply demands for shares that violate the constraint on short selling. The feasible optimum again requires a specialized portfolio and induces extreme dividend behavior in which one company pays no dividend and the other keeps no retained earnings. We return below to examine the characteristics of this "low-risk avoidance" equilibrium. Now however

<sup>16</sup>We indicate below the important circumstances under which the demands implied by this maximization would violate the "no short sale" constraints. This "limited-risk avoidance" case will be considered explicitly in Section IV.

we shall focus on the case in which  $A$  is nonsingular and the solution of equation (9) does not violate the other constraints on portfolio behavior.<sup>17</sup>

Solving equation (9) yields the households' share demand equation under the assumption that  $s_H \geq 0$  (i.e., that short selling would not be optimal):

$$(10) \quad s_H = \gamma^{-1} A^{-1} [a_H - R p].$$

Analogous share demand equations hold for the institutional investors:

$$(11) \quad s_I = \gamma^{-1} A^{-1} [a_I - R p].$$

The share demands differ only because  $a_H$  contains the tax variable ( $\theta > 0$ ) while in  $a_I$  the tax variable is implicitly zero.<sup>18</sup>

#### B. Price Functions and Optimal Dividends

By equating the share demands of (10) and (11) to the fixed-share supplies, we can solve for the market-clearing share prices that would correspond to any combination of dividend policies. Since the number of shares of each company was normalized to one, we have

$$(12) \quad s_H + s_I = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

or

$$(13) \quad \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \gamma^{-1} A^{-1} [a_H + a_I - 2R p].$$

Solving equation (13) for this price vector yields

$$(14) \quad p = \frac{1}{2R} \begin{bmatrix} R(2-\theta)d_1 + 2(1-d_1)r_1^e i \\ R(2-\theta)d_2 + 2(1-d_2)r_2^e i \end{bmatrix} - \frac{\gamma}{2R} \begin{bmatrix} (1-d_1)^2 \sigma_{11} + (1-d_1)(1-d_2)\sigma_{12} \\ (1-d_2)^2 \sigma_{22} + (1-d_1)(1-d_2)\sigma_{12} \end{bmatrix}.$$

<sup>17</sup>We later show that such equilibria can exist for plausible parameter values.

<sup>18</sup>If any of the nonnegativity constraints on  $s_H$  or  $s_I$  are binding, the optimum is no longer given by equations (10) and (11).

The price of each type of share is positively related to its own expected yield and negatively related to the variance of that yield and its covariance with the yield of the other type of stock.

We assume that each firm selects its dividend payout rate to maximize its share price and takes the dividend of the other firm as given.<sup>19</sup> The first-order condition for firm 1 is

$$(15) \quad \partial p_1 / \partial d_1 = 0 = [R(2-\theta) - 2r_1^e + \gamma[2(1-d_1)\sigma_{11} + (1-d_2)\sigma_{12}]] / 2R$$

and implies that the firm's optimal dividend rate ( $d_1^*$ ) satisfies

$$(16) \quad 1 - d_1^* = \frac{2(r_1^e - R) + \theta R}{2\gamma\sigma_{11}} - \frac{\sigma_{12}(1-d_2)}{2\sigma_{11}}.$$

Equation (16) describes the first firm's optimal reaction to the dividend policy of the second firm. Symmetrically we obtain the dividend policy reaction function of the second firm:

$$(17) \quad 1 - d_2^* = \frac{2(r_2^e - R) + \theta R}{2\gamma\sigma_{22}} - \frac{\sigma_{12}(1-d_1)}{2\sigma_{22}}.$$

If the returns to the investments by the two firms are not independent ( $\sigma_{12} \neq 0$ ), the optimal dividend policy of each firm depends on the dividend policy of the other firm. The two dividend policy functions can be solved simultaneously to obtain the equilibrium dividend policy of each firm:

$$(18) \quad \begin{bmatrix} 1 - d_1^* \\ 1 - d_2^* \end{bmatrix} = \frac{1}{1 - \frac{\sigma_{12}^2}{4\sigma_{11}\sigma_{22}}} \begin{bmatrix} \frac{2(r_1^e - R) + \theta R}{2\gamma\sigma_{11}} - \frac{\sigma_{12}}{2\sigma_{11}} \left[ \frac{2(r_2^e - R) + \theta R}{2\gamma\sigma_{22}} \right] \\ \frac{2(r_2^e - R) + \theta R}{2\gamma\sigma_{22}} - \frac{\sigma_{12}}{2\sigma_{22}} \left[ \frac{2(r_1^e - R) + \theta R}{2\gamma\sigma_{11}} \right] \end{bmatrix}.$$

<sup>19</sup>Section V develops a model with a large number of firms in which it is more natural to assume that each firm treats the dividends of other firms as parameters.

The stability of this solution may be easily verified by examining the reaction functions (16) and (17).

## II. Some Comparative Statics

It is immediately clear from equation (18) that each firm's optimal retained earnings depends positively on its own expected return and negatively on its own variance.<sup>20</sup> A higher expected yield makes it optimal to retain and invest more in the company while an increase in the uncertainty of that return makes the immediate payment of dividends more appealing.

If the returns of the two firms are positively correlated ( $\sigma_{12} > 0$ ), each firm's optimal retained earnings varies inversely with the attractiveness of investment in the other firm (i.e., with the other firm's expected yield and the inverse of its variance). Intuitively, when retained earnings in one firm are more attractive and therefore increase, the riskiness of retaining earnings in the other firm increases if the yields of the two firms are positively correlated.

The effect of an increase in the rate of tax on dividends is particularly interesting. For firm 1,

$$(19) \quad \frac{\partial d_1^*}{\partial \theta} = - \frac{1}{1 - \frac{\sigma_{12}^2}{4\sigma_{11}\sigma_{22}}} \cdot \frac{R}{2\gamma\sigma_{11}} \cdot \left[ 1 - \frac{\sigma_{12}}{2\sigma_{22}} \right].$$

The first two terms on the right-hand side are unambiguously positive. If the yields of the two firms are uncorrelated ( $\sigma_{12} = 0$ ), an increase in the tax rate on dividends necessarily reduces the firm's payout. However, when the yields are correlated the effect of the tax rate is ambiguous, that is, the sign of the final term in equation (19) can be either positive or negative. Since  $\sigma_{12}/\sigma_{22}$  is the

regression coefficient of the return for the first firm's investment on the return for the second firm's investment,<sup>21</sup> it could exceed 2 and make the final expression negative.

It is easy to understand why a strong covariance between the yields could produce the apparently counterintuitive result that an increase in the tax rate on dividends can actually raise a firm's optimal payout. Note first that an equation similar to (19) holds for firm 2:

$$(20) \quad \frac{\partial d_2^*}{\partial \theta} = - \frac{1}{1 - \frac{\sigma_{12}^2}{4\sigma_{11}\sigma_{22}}} \cdot \frac{R}{2\gamma\sigma_{22}} \cdot \left[ 1 - \frac{\sigma_{12}}{2\sigma_{11}} \right].$$

Adding these two expressions gives the effect of an increase in  $\theta$  on the total dividends of the two firms combined:

$$(21) \quad \frac{\partial (d_1^* + d_2^*)}{\partial \theta} = - \frac{1}{1 - \frac{\sigma_{12}^2}{4\sigma_{11}\sigma_{22}}} \cdot \frac{R}{2\gamma\sigma_{11}\sigma_{22}} [\sigma_{22} + \sigma_{11} - \sigma_{12}].$$

It is easy to show that this is unambiguously negative. This is clearly so if  $\sigma_{12} < 0$ . To see that this is also true when  $\sigma_{12} > 0$ , note that the variance of the difference  $r_1 - r_2$  is  $\sigma_{22} + \sigma_{11} - 2\sigma_{12}$ ; since this is a variance, it is necessarily positive, implying  $\sigma_{22} + \sigma_{11} > 2\sigma_{12}$  and therefore that  $\sigma_{22} + \sigma_{11} - \sigma_{12} > \sigma_{12} > 0$ . Thus an increase in the tax rate on dividends unambiguously reduces total dividends. The dividends of one of the firms may increase but not the dividends of both of them. The dividends of one firm will increase when the decrease in the dividends of the other is so large that, given the positive covariance between the returns, the greater risk associated

<sup>20</sup> Since  $\sigma_{12}^2/\sigma_{11}\sigma_{22}$  is the square of the correlation coefficient between the two yields and therefore necessarily less than unity, the common multiplier of both terms is positive.

<sup>21</sup> This regression coefficient is closely related to the *beta* of capital market theory, but refers here to the yields expressed as a return on physical capital rather than share value.

with retained earnings in the first firm outweighs the direct effect of the tax.

It is interesting to consider the magnitude of this sensitivity of the payout policy with respect to the tax parameter. Equation (18) can be used to calculate the elasticity of the aggregate retained earnings with respect to  $\theta$ . Although it is easy to obtain a general expression, the interpretation of the elasticity is clearer if we assume that the "excess yield" ( $r^e - R$ ) is the same for both assets.<sup>22</sup> With this assumption, equation (18) implies the elasticity

$$(22) \quad \frac{\theta}{2 - d_1^* - d_2^*} \cdot \frac{\partial(2 - d_1^* - d_2^*)}{\partial \theta} = \frac{\theta R}{2(r^e - R) + \theta R}.$$

In the special case in which the expected yield is equal to the yield on the riskless asset (i.e.,  $r^e = R$ ), there are retained earnings only because of the tax effect and the elasticity of the retained earnings with respect to the dividend tax rate is unity. When there is a positive expected excess return on retained earnings, the tax effect is less important and the elasticity is less than one.<sup>23</sup>

### III. Characteristics of the Symmetric Equilibrium

The special case in which the two firms have equal expected yields and equal variances is particularly interesting to analyze. Together with the assumptions that we have made about the similarity of the two types of investors, this assumption about the firm implies that the only essential source of difference in the model is in the different tax

treatments of households and institutions. We commented in the introduction, and show formally in Section IV below, that when the advantage of diversification is small (i.e., low risk or high correlation), this difference in taxation leads to specialization of ownership and corner solutions for the firms' dividend policies; that is, the firm that remains in business pays no dividend. We now examine the characteristics of the equilibrium in the case in which there is sufficient risk and opportunity for diversification and show that in this case both firms do pay dividends. *The opportunity for advantageous diversification by investors induces positive dividends by firms.*

With  $r_1^e = r_2^e$  and  $\sigma_{11} = \sigma_{22}$ , equation (18) shows immediately that  $d_1^* = d_2^*$ , that is, both firms have the same optimal dividend. In contrast to the "no-diversification" case in which the dividend policies are at opposite extremes, advantageous diversification produces identical dividend policies. This common dividend policy satisfies

$$(23) \quad 1 - d^* = (2(r^e - R) + \theta R) / \gamma \sigma (2 + \rho),$$

where  $r^e$  is the common expected yield,  $\sigma$  is the common variance, and  $\rho$  is the correlation between the yields.<sup>24</sup>

Note first that  $\theta = 0$  and  $r^e = R$  together imply  $d^* = 1$ ; when there is no tax on dividends and no "excess return" on funds retained in the firm, all profits will be paid out. The economic reason for this is clear: with no tax or yield incentive for retention, full payout avoids the risk of retained earnings without any loss in after-tax yield.

A small tax on dividends clearly makes  $1 - d^* > 0$  and therefore  $d^* < 1$ , that is, both firms pay out some but not all of their profits as dividends. A positive but partial dividend payout is clearly optimal despite a tax that discriminates against dividends. Of course, a large enough value of  $\theta$  can make  $1 - d^* \geq 1$  and therefore imply  $d^* = 0$ ; when the tax discrimination against dividends is

<sup>22</sup>When the excess returns differ for the two firms,  $r^e - R$  is replaced by a weighted average including the variances and covariances of the yields.

<sup>23</sup>In an early empirical study of the effect of taxes on the dividend policy of British firms, Feldstein (1970) estimated that the elasticity of the dividend rate with respect to the inverse of  $\theta$  was 0.9. Since dividends were about two-thirds of retained earnings in that sample period, the estimated elasticity of 0.9 corresponds to an elasticity of retained earnings with respect to  $\theta$  of approximately 0.6, and is therefore quite compatible with equation (22).

<sup>24</sup>With  $\sigma_{11} = \sigma_{22}$ ,  $\rho = \sigma_{12} / \sigma_{22} = \sigma_{12} / \sigma_{11}$ . Equation (23) follows directly from (18) when it is noted that the common multiplier in (18) is the inverse of  $1 - (\rho/2)^2$  and that  $1 - (\sigma_{12} / 2\sigma_{11}) = 1 - (\rho/2)$ ; the ratio of these two is the inverse of  $1 + (\rho/2)$ .

strong enough, no dividends will be paid. Note that the excess return on retained earnings affects the optimal dividends in the same way as the dividend tax. Starting at  $\theta = 0$  and  $r^e = R$ , a small increase in  $r^e$  will cause positive but partial dividend payout while a large enough excess return on retained earnings will cause all dividends to stop.<sup>25</sup>

Consider next the price per share that prevails in this case when both firms adopt the optimal dividend policy. This share price is the value that investors place on the initial dollar of available profits inside the firm.<sup>26</sup> Since dollars retained in the firms have equal expected yields and equal variances, their share prices must also be equal. Equation (14) confirms this and shows that the common price is

$$(24) \quad p = R^{-1} \left[ (1 - \theta/2) dR + (1 - d) r^e - \gamma(1 - d)^2 \sigma(1 + \rho)/2 \right].$$

The three terms on the right-hand side of (24) show that the price depends on the

net-of-tax value of the current dividend  $[(1 - \theta/2)d]$ , the expected present value of the retained earnings  $[(1 - d)r^e R^{-1}]$ , and the offset for the risk associated with the retained earnings  $[\gamma(1 - d)^2 \sigma(1 + \rho)] R^{-1}$ . Substituting the optimal value of the dividend payout rate from equation (23) and rearranging terms yields

$$(25) \quad p = 1 - \theta/2 + \gamma \sigma (1 - d^*)^2 / 2R$$

or

$$(26) \quad p = 1 - \frac{\theta}{2} + \frac{[r^e - R + R\theta/2]^2}{2R\gamma\sigma(1 + \rho/2)^2}.$$

Since half of the shareholders pay tax at rate  $\theta$  while half pay no tax, the average tax rate is  $\theta/2$  and  $1 - \theta/2$  is the net-of-tax income per dollar of dividends. Equation (25) shows that when it is optimal to pay out all profits as dividends ( $d^* = 1$ ), the share price equals the net-of-tax value of the dividend.<sup>27</sup> More generally, when the firms retain some of their earnings, the price per share exceeds the net-of-tax amount that could be distributed. This is shown clearly in equation (25). The equivalent expression in equation (26) indicates why this is so. Since it is optimal for a firm to retain some of its earnings when the returns inside the firm exceed their opportunity cost or when there is a tax penalty on dividends, either of these reasons to limit dividends causes an increase in the share price vis-à-vis the price that would prevail if  $d^* = 1$ . This is seen explicitly in equation (26). To the extent that there is an excess return on retained earnings ( $r^e > R$ ), or that the average tax rate on dividends is positive ( $\theta/2 > 0$ ), the price exceeds the net amount that could be distributed. An increase in risk aversion ( $\gamma$ ) or in the riskiness of retained earnings ( $\sigma(1 + \rho/2)$ ) decreases the magnitude of this premium.

It is interesting to note that the price per dollar of earnings inside the firm may be less

<sup>25</sup>It is tempting to ask what happens as  $\rho$  tends to unity. When  $\rho = 1$ , there is no opportunity for diversification. The economics implies that in this case there will be specialization of ownership and therefore of dividend policy. This *cannot* be seen by setting  $\rho = 1$  in equation (23) because (23) does not hold when  $\rho = 1$ . When  $\rho = 1$ , the matrix  $A$  of equation (9) is singular and the share demand equations ((10) and (11)) from which (23) is derived do not hold.

<sup>26</sup>There is an extensive literature on this value, which is sometimes referred to as "Tobin's  $q$ ." It has been common to assume that the equilibrium value of  $q$  is one, an assumption that we accepted in our paper with Sheshinski. Auerbach, Bradford, and King analyze a model without uncertainty and with only taxable shareholders; they conclude that if firms are paying positive but partial dividends, the share price must equal  $1 - \theta$ , i.e., a dollar of profits inside the firm must be valued at the amount that can be paid *net of tax* to the shareholder. (Their analysis also allows for a tax on capital gains; which also influences the share price; in the absence of this tax their share price formula reduces to  $1 - \theta$ .) Studies using the capital asset pricing model to measure the value of a *marginal* dollar inside the firm produce estimates that vary substantially over time with an average that is in the range of unity or somewhat less; see Gordon and Bradford and the studies that they cite. Green shows that changes in share prices on their ex-dividend days cannot be used to estimate the value of a marginal dollar of funds inside the firm.

<sup>27</sup>This special case thus corresponds to the Auerbach-Bradford-King share-price equation extended to the case of heterogeneous taxpayers. It holds however only when all profits are paid as dividends by both firms.

than, equal to, or greater than unity. When  $d^* = 1$ , the price is clearly less than 1. A high value of excess return can of course produce a share value greater than one. But even if  $r^e = R$ , the price lies between  $1 - \theta/2$  and 1.<sup>28</sup>

Our discussion in this section has implicitly assumed that both types of investors hold both assets in an optimal portfolio. It can be demonstrated that this is in fact true unless the product of the risk-aversion parameter ( $\gamma$ ), the common variance ( $\sigma_{11} = \sigma_{22}$ ), and the tax rate ( $\theta$ ) are relatively high. For a high enough value of  $\theta\gamma\sigma_{11}$ , the taxable individuals will wish to hold only the riskless asset with yield  $R$ . In this case, the shares are held only by the tax-free institutions. But if risk aversion and risk are not too high, individuals as well as institutions will want to hold positive amounts of the shares of both firms.

#### A. A Numerical Example

To conclude this analysis of the case in which the opportunity for advantageous diversification causes nonspecialization and positive but partial dividend payout, it is useful to present a numerical example in which these properties hold. Consider the case in which the expected return on investment in both firms is  $r^e = 1.3$  and the correlation between the return is  $\rho = 0.5$ . Let the tax rate be  $\theta = 0.5$  and the riskless yield on the alternative asset be  $R = 1.1$ . The common variance of the returns does not matter as such, only the product of the variance and the risk-aversion coefficient ( $\gamma\sigma$ ). The dividend payout rate ( $d$ ) and the combined risk parameter ( $\gamma\sigma$ ) must satisfy the dividend payout condition (equation (23)) and the condition that the demand for shares by households and institutions (given by equations (10) and (11)) together equal unity for each firm and separately do not violate the condition that investors may not sell short.

The symmetry of the current problem implies that each type of investor will hold equal amounts of both types of shares. These conditions are satisfied if the dividend payout rate is  $d = 0.8$  and the risk parameter is  $\gamma\sigma = 1.87$ . Equation (25) implies that the corresponding price per share is  $p = 0.78$ .

#### IV. The Segmented Market Equilibrium

We have been analyzing the case in which firms are identical but in which there is enough opportunity for advantageous diversification to cause investors to hold mixed portfolios. Firms pay out positive dividends in a value-maximizing equilibrium. Qualitatively, these results are not surprising. It is, however, somewhat odd that the equilibrium of our model in the symmetric case is itself symmetric: both firms choose the same dividend payout rate and each investor holds an equal share in the two firms. The conflict between diversification and tax avoidance is completely resolved in favor of the former. One might have thought that the firms would "locate" at different points in the dividend spectrum attracting a different clientele, one more heavily taxed on average than the other, and that investors would accept this incomplete diversification in equilibrium in order to reap the tax advantages.

At present we do not know whether this striking symmetry property is the result of the mean-variance utility, the "two-class" model of investors, or whether it is a phenomenon of more fundamental generality.

In this section we will show that this symmetric equilibrium, which is unique whenever investors are holding shares of both firms, coexists with asymmetric "locational" equilibria when the nonnegativity conditions for portfolios are binding. Such a situation arises when there is little variance in yields or a high correlation between the two firms so that diversification is of only limited benefit.

The phenomenon of asymmetric, segmented market equilibrium is seen most clearly in the extreme case of certainty:  $\sigma_{11} = \sigma_{22} = 0$ . This lack of risk implies that each investor values shares at the present value of their payouts, net of taxes. For either firm, one dollar paid as dividends is worth  $(1 - \theta)R$

<sup>28</sup>Clearly when  $r^e = R$  and  $d^* = 0$ , the value of the firm is the discounted expected value of the subsequent payout ( $r^e/R = 1$ ) minus any adjustment for risk. When  $r^e = R$  but  $d^* > 0$ ,  $p$  lies between this upper bound and  $1 - \theta/2$ .

to households and  $R$  to institutions; while one dollar of retained earnings is worth  $r$  to both types of investors.

Consider the case in which  $R > r^e > R(1 - \theta)$ , that is, in which funds inside the firm have a lower yield than outside the firm ( $R > r^e$ ), but are worth more than funds outside the firm if a dividend tax has to be paid ( $r^e > (1 - \theta)R$ ).<sup>29</sup> In this case, the untaxed institutional investor prefers immediate payout ( $d = 1$ ) because the value of the dividend ( $R$ ) exceeds the expected value of the funds left in the company ( $r^e$ ). In contrast, the taxed household investor prefers no dividend payout ( $d = 0$ ), because the value of the net-of-tax dividend ( $(1 - \theta)R$ ) is less than the expected value of the funds left in the company ( $r^e$ ). The market will accommodate this conflict of preferences by specialization of ownership and dividend policies.

Let us examine the equilibrium prices that would lead to  $d_1 = 0$ ,  $d_2 = 1$ , with portfolios  $s_{H1} = 1$ ,  $s_{H2} = 0$ ,  $s_{I1} = 0$ ,  $s_{I2} = 1$ . First, it is clear that, unless the initial ownership of shares gives the two classes equal portfolio wealth, the equilibrium prices of the two firms may not be equal. This is not incompatible with the value-maximizing assumption because the firm cannot achieve the other's value by mimicking its dividend policy. Both values will change in this process.

We will show that the equilibrium prices are given by

$$(27) \quad p_1 = R^{-1}r^e; \quad p_2 \in [(1 - \theta), 1],$$

where the precise value of  $p_2$  in this interval is determined in such a way that the portfolios described above are compatible with the budget equation (1). For households to hold shares of firm 1 in positive quantity, we need  $p_1 \leq R^{-1}r^e$  and if they don't hold firm 2, then  $p_2 \geq 1 - \theta$ . Similarly, the implications that can be derived from institutions' portfolios are  $p_2 \leq 1$ ,  $p_1 \geq R^{-1}r^e$ . Combining these we see that (27) is required.

<sup>29</sup>In the alternative case of  $r^e > R$ , both investors will prefer to have no dividends and both firms will therefore choose  $d = 0$ . The firms behave identically and there is no market segmentation.

To verify that these prices are indeed equilibria, it is necessary to see what changes would be induced by different dividend policies. This problem is a little curious in that even if dividends were to vary, the same prices and the same portfolios could still persist. Thus the equilibrium sustained by extreme dividend policies is compatible with value maximization only in the sense that firms are indifferent to these choices.

It is of interest to note that the symmetric equilibrium  $d_1 = d_2 = 0$  and  $p_1 = p_2 = R^{-1}r^e$  is also an equilibrium here.<sup>30</sup> The paradox of symmetric vs. segmented equilibria is resolved by noting that the latter are produced when the nonnegativity constraints for portfolios are binding.

Moreover one can observe that since no taxes are actually collected in either of these cases, the consumption patterns, and hence welfare considerations, are identical.

The results of the riskless case can be extended to the case of small variance or high correlation without changing the essential conclusion. In such cases, the share demands implied by equations (10) and (11) would violate the no short-selling constraint. The constrained optimum would involve a corner solution in which ownership is specialized. The dividend policy of each company would then be adjusted to the tax situation of its homogeneous group of shareholders. The lack of such homogeneity and the presence of dividends for the majority of major publicly owned companies suggest that the opportunities for advantageous diversification are sufficient to prevent shareholder specialization.<sup>31</sup>

## V. Conclusion

This paper has provided a simple model of market equilibrium to explain why firms that maximize the value of their shares pay dividends even though the funds could instead

<sup>30</sup>Any share ownership will sustain this, and individuals will be indifferent.

<sup>31</sup>Other possibilities include a nonhomogeneity of beliefs which are not perfectly correlated with tax status, locked-in investors due to the taxation of capital gains on realization, or intertemporal considerations which are of practical importance but are difficult to model.

be retained and subsequently distributed to shareholders in a way that would allow them to be taxed more favorably as capital gains. Our explanation does not rely on any asymmetry of information or divergence of interests between management and shareholders. The heterogeneity of tax rates and the existence of uncertainty and of risk aversion are explicitly recognized. Indeed, it is the combination of the conflicting preferences of shareholders in different tax brackets and their desire for portfolio diversification in the face of uncertainty that together cause all firms to pay dividends in our model.

The model that we have used should be extended in several directions in order to provide a more realistic framework for analysis. The most important extension would be to an economy with many firms. It can be shown that such an extension preserves the main results of the two-firm model, including all of the characteristics of the symmetric equilibria, if the variance of each firm's return grows with the number of firms in the economy.<sup>32</sup> In contrast, if the variance of each firm remains constant, an increased number of firms will cause a segmented equilibrium in which taxable shareholders invest in one diversified portfolio of firms and non-taxable shareholders invest in a different portfolio. This occurs because each investor's portfolio becomes progressively less risky as the number of firms increase, inducing the investor to concentrate on the tax advantages of a specialized portfolio even though that requires some loss of diversification.

We believe that an alternative and more natural generalization that preserves the non-segmented market equilibrium is to recognize that, within each tax class, investors have heterogeneous expectations about individual firms. This implies that each firm is subjectively unique, and that both high- and low-tax investors will want to invest in all firms. We hope to present an explicit analysis of this multifirm case in a subsequent paper.

Another worthwhile extension of the present model would be to recognize that both corporations and portfolio investors can also borrow and that corporations as well as in-

vestors can earn the risk-free return. Such a model would have to introduce a layer of corporate taxation if misleading results are to be avoided. Expanding the firm's financial behavior in this way would weaken the link between dividends and real corporate investment that is in the present model. We believe, however, that such a link between dividend policy and real corporate investment would persist, contrary to the Modigliani-Miller theorem or the complete separation of investment and financial decisions.

An explicit multiperiod analysis with growing capital stocks should also be developed. The relationship between each firm's rate of investment and its equilibrium rate of return can be analyzed within this extended framework.

The present study indicates that the existing tax treatment of dividends distorts corporate financial decisions and may cause a misallocation of total investment. It will be important to see whether these adverse effects remain in the more general analytic framework.

## REFERENCES

- Auerbach, Alan, "Wealth Maximization and the Cost of Capital," *Quarterly Journal of Economics*, August 1979, 93, 433-46.
- Bhattacharya, Sudipto, "Imperfect Information, Dividend Policy, and 'The Bird in the Hand' Fallacy," *Bell Journal of Economics*, Spring 1979, 10, 259-70.
- Bradford, David, "The Incidence and Allocation Effect of a Tax on Corporate Distributions," Working Paper No. 349, National Bureau of Economics, May 1979.
- Feenberg, Daniel, "Does the Investment Interest Limitation Explain the Existence of Dividends?," *Journal of Financial Economics*, Summer 1981, 9, 265-69.
- Feldstein, Martin, "Corporate Taxation and Dividend Behavior," *Review of Economic Studies*, January 1970, 37, 57-72.
- \_\_\_\_\_, Green, Jerry and Sheshinski, Eytan "Corporate Financial Policy and Taxation in a Growing Economy," *Quarterly Journal of Economics*, August 1979, 93, 411-32.
- \_\_\_\_\_, and Slemrod, Joel, "Personal Taxation, Portfolio Choice and the Effect of the

<sup>32</sup>This case is developed in Sections 5 and 6 of our earlier NBER version of the current paper.

- Corporation Income Tax," *Journal of Political Economy*, October 1980, 88, 854-66.
- Gordon, Roger and Bradford, David, "Stock Market Valuation of Dividends: Links to the Theory of the Firm and Empirical Estimates," mimeo., May 1979.
- \_\_\_\_\_ and Malkiel, Burton "Corporate Financial Structure," paper presented at Brookings Conference on Economic Effects of Federal Taxes, October 18-19, 1979.
- Green, Jerry, "Taxes and the Ex-Dividend Pay Behavior of Common Stock Prices," Discussion Paper No. 772, Harvard Institute of Economic Research, June 1980.
- King, Mervyn, *Public Policy and the Corporation*, London: Chapman and Hall, Ltd., 1977.
- Miller, Merton and Scholes, Myron, "Dividends and Taxes," *Journal of Financial Economics*, December 1978, 64, 333-64.
- Modigliani, Franco and Miller, Merton, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *American Economic Review*, June 1958, 48, 261-97.
- Ross, Stephen, "The Determination of Financial Structures: The Incentive Signalling Approach," *Bell Journal of Economics*, Spring 1977, 8, 23-40.

# Let's Take the Con out of Econometrics

By EDWARD E. LEAMER\*

Econometricians would like to project the image of agricultural experimenters who divide a farm into a set of smaller plots of land and who select randomly the level of fertilizer to be used on each plot. If some plots are assigned a certain amount of fertilizer while others are assigned none, then the difference between the mean yield of the fertilized plots and the mean yield of the unfertilized plots is a measure of the effect of fertilizer on agricultural yields. The econometrician's humble job is only to determine if that difference is large enough to suggest a real effect of fertilizer, or is so small that it is more likely due to random variation.

This image of the applied econometrician's art is grossly misleading. I would like to suggest a more accurate one. The applied econometrician is like a farmer who notices that the yield is somewhat higher under trees where birds roost, and he uses this as evidence that bird droppings increase yields. However, when he presents this finding at the annual meeting of the American Ecological Association, another farmer in the audience objects that he used the same data but came up with the conclusion that moderate amounts of shade increase yields. A bright chap in the back of the room then observes that these two hypotheses are indistinguishable, given the available data. He mentions the phrase "identification problem," which, though no one knows quite what he means, is said with such authority that it is totally convincing. The meeting reconvenes in the halls and in the bars, with heated discussion whether this is the kind of work that merits promotion from Associate to Full Farmer; the Luminists strongly opposed to promotion and the Aviophiles equally strong in favor.

One should not jump to the conclusion that there is necessarily a substantive difference between drawing inferences from experimental as opposed to nonexperimental data. The images I have drawn are deliberately prejudicial. First, we had the experimental scientist with hair neatly combed, wide eyes peering out of horn-rimmed glasses, a white coat, and an electronic calculator for generating the random assignment of fertilizer treatment to plots of land. This seems to contrast sharply with the nonexperimental farmer with overalls, unkempt hair, and bird droppings on his boots. Another image, drawn by Orcutt, is even more damaging: "Doing econometrics is like trying to learn the laws of electricity by playing the radio." However, we need not now submit to the tyranny of images, as many of us have in the past.

## I. Is Randomization Essential?

What is the real difference between these two settings? Randomization seems to be the answer. In the experimental setting, the fertilizer treatment is "randomly" assigned to plots of land, whereas in the other case nature did the assignment. Now it is the tyranny of words that we must resist. "Random" does not mean adequately mixed in *every* sample. It only means that on the average, the fertilizer treatments are adequately mixed. Randomization implies that the least squares estimator is "unbiased," but that definitely does not mean that for each sample the estimate is correct. Sometimes the estimate is too high, sometimes too low. I am reminded of the lawyer who remarked that "when I was a young man I lost many cases that I should have won, but when I grew older I won many that I should have lost, so on the average justice was done."

In particular, it is possible for the randomized assignment to lead to exactly the same allocation as the nonrandom assignment,

\*Professor of economics, University of California-Los Angeles. This paper was a public lecture presented at the University of Toronto, January 1982. I acknowledge partial support by NSF grant SOC78-09479.

namely, with treated plots of land all being under trees and with nontreated plots of land all being away from trees. I submit that, if this is the outcome of the randomization, then the randomized experiment and the nonrandomized experiment are exactly the same. Many econometricians would insist that there is a difference, because the randomized experiment generates "unbiased" estimates. But all this means is that, if this particular experiment yields a gross overestimate, some other experiment yields a gross underestimate.

Randomization thus does not assure that each and every experiment is "adequately mixed," but randomization does make "adequate mixing" probable. In order to make clear what I believe to be the true value of randomization, let me refer to the model

$$(1) \quad Y_i = \alpha + \beta F_i + \gamma L_i + U_i,$$

where  $Y_i$  is the yield of plot  $i$ ;  $F_i$  is the fertilizer assigned to plot  $i$ ;  $L_i$  is the light falling on plot  $i$ ;  $U_i$  is the unspecified influence on the yield of plot  $i$ , and where  $\beta$ , the fertilizer effect, is the object of the inferential exercise. We may suppose to begin the argument that the light level is expensive to measure and that it is decided to base an estimate of  $\beta$  initially only on measurement of  $Y_i$  and  $F_i$ . We may assume also that the natural experiment produces values for  $F_i$ ,  $L_i$ , and  $U_i$  with expected values  $E(U_i|F_i) = 0$  and  $E(L_i|F_i) = r_0 + r_1 F_i$ . In the more familiar parlance, it is assumed that the fertilizer level and the residual effects are uncorrelated, but the fertilizer level and the light level are possibly correlated. As every beginning econometrics student knows, if you omit from a model a variable which is correlated with included variables, bad things happen. These bad things are revealed to the econometrician by computing the conditional mean of  $Y$  given  $F$  but not  $L$ :

$$\begin{aligned} (2) \quad E(Y|F) &= \alpha + \beta F + \gamma E(L|F) \\ &= \alpha + \beta F + \gamma(r_0 + r_1 F) \\ &= (\alpha + \alpha^*) + (\beta + \beta^*)F, \end{aligned}$$

where  $\alpha^* = \gamma r_0$  and  $\beta^* = \gamma r_1$ . The linear regression of  $Y$  on  $F$  provides estimates of the parameters of the conditional distribution of  $Y$  given  $F$ , and in this case the regression coefficients are estimates not of  $\alpha$  and  $\beta$ , but rather of  $\alpha + \alpha^*$  and  $\beta + \beta^*$ . The parameters  $\alpha^*$  and  $\beta^*$  measure the bias in the least squares estimates. This bias could be due to left-out variables, or to measurement errors in  $F$ , or to simultaneity.

When observing a nonexperiment, the bias parameters  $\alpha^*$  and  $\beta^*$  can be thought to be small, but they cannot sensibly be treated as exact zeroes. The notion that the bias parameters are small can be captured by the assumption that  $\alpha^*$  and  $\beta^*$  are drawn from a normal distribution with zero means and covariance matrix  $M$ . The model can then be written as  $Y = \alpha + \beta F + \varepsilon$ , where  $\varepsilon$  is the sum of three random variables:  $U + \alpha^* + \beta^* F$ . Because the error term  $\varepsilon$  is not spherical, the proper way to estimate  $\alpha$  and  $\beta$  is generalized least squares. My 1974 article demonstrates that if  $(a, b)$  represent the least squares estimates of  $(\alpha, \beta)$ , then the generalized least squares estimates  $(\hat{\alpha}, \hat{\beta})$  are also equal to  $(a, b)$ :

$$(3) \quad \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix},$$

and if  $S$  represents the sample covariance matrix for the least squares estimates, then the sample covariance matrix for  $(\hat{\alpha}, \hat{\beta})$  is

$$(4) \quad \text{Var}(\hat{\alpha}, \hat{\beta}) = S + M,$$

where  $M$  is the covariance matrix of  $(\alpha^*, \beta^*)$ .

The meaning of equation (3) is that unless one knows the direction of the bias, the possibility of bias does not call for any adjustment to the estimates. The possibility of bias does require an adjustment to the covariance matrix (4). The uncertainty is composed of two parts: the usual sampling uncertainty  $S$  plus the misspecification uncertainty  $M$ . As sample size grows, the sampling uncertainty  $S$  ever decreases, but the misspecification uncertainty  $M$  remains ever constant. The misspecification matrix  $M$  that we must add to the least squares variance

matrix is just the (prior) variance of the bias coefficients ( $\alpha^*$ ,  $\beta^*$ ). If this variance matrix is small, the least squares bias is likely to be small. If  $M$  is large, it is correspondingly probable that ( $\alpha^*$ ,  $\beta^*$ ) is large.

It would be a remarkable bootstrap if we could determine the extent of the misspecification from the data. The data in fact contain no information about the size of the bias, a point which is revealed by studying the likelihood function. The misspecification matrix  $M$  is therefore a pure prior concept. One must decide independent of the data how good the nonexperiment is.

The formal difference between a randomized experiment and a natural experiment is measured by the matrix  $M$ . If the treatment is randomized, the bias parameters ( $\alpha^*$ ,  $\beta^*$ ) are exactly zero, or, equivalently, the matrix  $M$  is a zero matrix. If  $M$  is zero, the least squares estimates are consistent. If  $M$  is not zero, as in the natural experiment, there remains a fixed amount of specification uncertainty, independent of sample size.

There is therefore a sharp difference between inference from randomized experiments and inference from natural experiments. This seems to draw a sharp distinction between economics where randomized experiments are rare and "science" where experiments are routinely done. But the fact of the matter is that no one has ever designed an experiment that is free of bias, and no one can. As it turns out, the technician who was assigning fertilizer levels to plots of land, took his calculator into the fields, and when he was out in the sun, the calculator got heated up and generated large "random" numbers, which the technician took to mean no fertilizer; and when he stood under the shade of the trees, his cool calculator produced small numbers, and these plots received fertilizer.

You may object that this story is rather fanciful, but I need only make you think it is possible, to force you to set  $M \neq 0$ . Or if you think a computer can really produce random numbers (calculated by a mathematical formula and therefore perfectly predictable!), I will bring up mismeasurement of the fertilizer level, or human error in carrying out the computer instructions. Thus, the attempt to

randomize and the attempt to measure accurately ensures that  $M$  is small, but not zero, and the difference between scientific experiments and natural experiments is difference in degree, but not in kind. Admittedly however, the misspecification uncertainty in many experimental settings may be so small that it is well approximated by zero. This can very rarely be said in nonexperimental settings.

Examples may be ultimately convincing. There is a great deal of empirical knowledge in the science of astronomy, yet there are no experiments. Medical knowledge is another good example. I was struck by a headline in the January 5, 1982 *New York Times*: "Life Saving Benefits of Low-Cholesterol Diet Affirmed in *Rigorous Study*." The article describes a randomized experiment with a control group and a treated group. "Rigorous" is therefore interpreted as "randomized." As a matter of fact, there was a great deal of evidence suggesting a link between heart disease and diet before any experiments were performed on humans. There were cross-cultural comparisons and there were animal studies. Actually, the only reason for performing the randomized experiment was that someone believed there was pretty clear nonexperimental evidence to begin with. The nonexperimental evidence was, of course, inconclusive, which in my language means that the misspecification uncertainty  $M$  remained uncomfortably large. The fact that the Japanese have both less incidence of heart disease and also diets lower in cholesterol compared to Americans is not convincing evidence, because there are so many other factors that remain unaccounted for. The fact that pigs on a high cholesterol diet develop occluded arteries is also not convincing, because the similarity in physiology in pigs and humans can be questioned.

When the sampling uncertainty  $S$  gets small compared to the misspecification uncertainty  $M$ , it is time to look for other forms of evidence, experiments or nonexperiments. Suppose I am interested in measuring the width of a coin, and I provide rulers to a room of volunteers. After each volunteer has reported a measurement, I compute the mean and standard deviation, and I conclude that

the coin has width 1.325 millimeters with a standard error of .013. Since this amount of uncertainty is not to my liking, I propose to find three other rooms full of volunteers, thereby multiplying the sample size by four, and dividing the standard error in half. That is a silly way to get a more accurate measurement, because I have already reached the point where the sampling uncertainty  $S$  is very small compared with the misspecification uncertainty  $M$ . If I want to increase the true accuracy of my estimate, it is time for me to consider using a micrometer. So too in the case of diet and heart disease. Medical researchers had more or less exhausted the vein of nonexperimental evidence, and it became time to switch to the more expensive but richer vein of experimental evidence.

In economics, too, we are switching to experimental evidence. There are the laboratory experiments of Charles Plott and Vernon Smith (1978) and Smith (1980), and there are the field experiments such as the Seattle/Denver income maintenance experiment. Another way to limit the misspecification error  $M$  is to gather different kinds of nonexperiments. Formally speaking, we will say that experiment 1 is qualitatively different from experiment 2 if the bias parameters  $(\alpha_1^*, \beta_1^*)$  are distributed independently of the bias parameters  $(\alpha_2^*, \beta_2^*)$ . In that event, simple averaging of the data from the two experiments yields average bias parameters  $(\alpha_1^* + \alpha_2^*, \beta_1^* + \beta_2^*)/2$  with misspecification variance matrix  $M/2$ , half as large as the (common) individual variances. Milton Friedman's study of the permanent income hypothesis is the best example of this that I know. Other examples are hard to come by. I believe we need to put much more effort into identifying qualitatively different and convincing kinds of evidence.

Parenthetically, I note that traditional econometric theory, which does not admit experimental bias, as a consequence also admits no "hard core" propositions. Demand curves can be shown to be positively sloped. Utility can be shown not to be maximized. Econometric evidence of a positively sloped demand curve would, as a matter of fact, be routinely explained in terms of simultaneity bias. If utility seems not to have been maxi-

mized, it is only that the econometrician has misspecified the utility function. The misspecification matrix  $M$  thus forms Imre Lakatos' "protective belt" which protects certain hard core propositions from falsification.

## II. Is Control Essential?

The experimental scientist who notices that the fertilizer treatment is correlated with the light level can correct his experimental design. He can control the light level, or he can allocate the fertilizer treatment in such a way that the fertilizer level and the light level are not perfectly correlated.

The nonexperimental scientist by definition cannot control the levels of extraneous influences such as light. But he can control for the variable light level by including light in the estimating equation. Provided nature does not select values for light and values for fertilizer levels that are perfectly correlated, the effect of fertilizer on yields can be estimated with a multiple regression. The collinearity in naturally selected treatment variables may mean that the data evidence is weak, but it does not invalidate in any way the usual least squares estimates. Here, again, there is no essential difference between experimental and nonexperimental inference.

## III. Are the Degrees of Freedom Inadequate with Nonexperimental Data?

As a substitute for experimental control, the nonexperimental researcher is obligated to include in the regression equation all variables that might have an important effect. The NBER data banks contain time-series data on 2,000 macroeconomic variables. A model explaining gross national product in terms of all these variables would face a severe degrees-of-freedom deficit since the number of annual observations is less than thirty. Though the number of observations of any phenomenon is clearly limited, the number of explanatory variables is logically unlimited. If a polynomial could have a degree as high as  $k$ , it would usually be admitted that the degree could be  $k+1$  as well. A theory that allows  $k$  lagged explanatory vari-

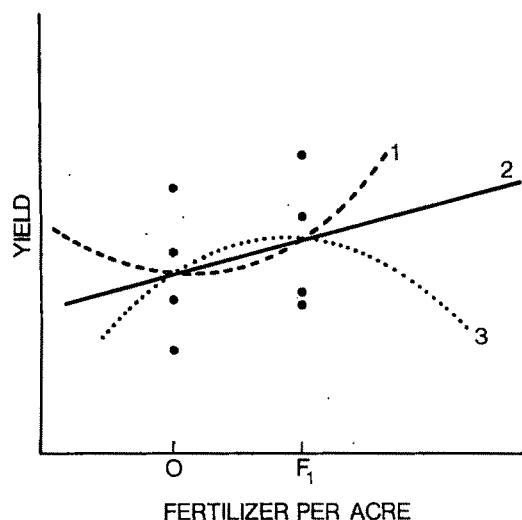


FIGURE 1. HYPOTHETICAL DATA AND THREE ESTIMATED QUADRATIC FUNCTIONS

ables would ordinarily allow  $k+1$ . If the level of money might affect *GNP*, then why not the number of presidential sneezes, or the size of the polar ice cap?

The number of explanatory variables is unlimited in a nonexperimental setting, but it is also unlimited in an experimental setting. Consider again the fertilizer example in which the farmer randomly decides either to apply  $F_1$  pounds of fertilizer per acre or zero pounds, and obtains the data illustrated in Figure 1. These data admit the inference that fertilizer level  $F_1$  produces higher yields than no fertilizer. But the farmer is interested in selecting the fertilizer level that maximizes profits. If it is hypothesized that yield is a linear function of the fertilizer intensity  $Y = \alpha + \beta F + U$ , then profits are

$$\text{Profits} = pA(\alpha + \beta F + U) - p_F AF,$$

where  $A$  is total acreage,  $p$  is the product price, and  $p_F$  is the price per pound of fertilizer. This profit function is linear in  $F$  with slope  $A(\beta p - p_F)$ . The farmer maximizes profits therefore by using no fertilizer if the price of fertilizer is high,  $\beta p < p_F$ , and using an unlimited amount of fertilizer if the price is low,  $\beta p > p_F$ . It is to be expected that you will find this answer unacceptable for one of

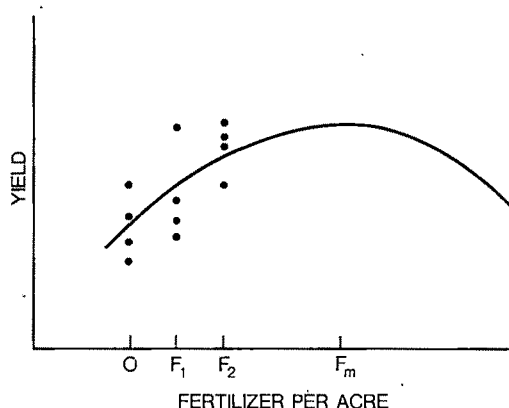


FIGURE 2. HYPOTHETICAL DATA AND ESTIMATED QUADRATIC FUNCTION

several reasons:

1) When the farmer tries to buy an unlimited amount of fertilizer, he will drive up its price, and the problem should be reformulated to make  $p_F$  a function of  $F$ .

2) Uncertainty in the fertilizer effect  $\beta$  causes uncertainty in profits,  $\text{Variance}(\text{profits}) = p^2 A^2 F^2 \text{Var}(\beta)$ , and risk aversion will limit the level of fertilizer applied.

3) The yield function is nonlinear.

Economic theorists doubtless find reasons 1) and 2) compelling, but I suspect that the real reason farmers don't use huge amounts of fertilizer is that the marginal increase in the yield eventually decreases. Plants don't grow in fertilizer alone.

So let us suppose that yield is a quadratic function of fertilizer intensity,  $Y = \alpha + \beta_1 F + \beta_2 F^2 + U$ , and suppose we have only the data illustrated in Figure 1. Unfortunately, there are an infinite number of quadratic functions all of which fit the data equally well, three of which are drawn. If there were no other information available, we could conclude only that the yield is higher at  $F_1$  than at zero. Formally speaking, there is an identification problem, which can be solved by altering the experimental design. The yield must be observed at a third point, as in Figure 2, where I have drawn the least squares estimated quadratic function and have indicated the fertilizer intensity  $F_m$  that maximizes the yield. I expect that most people would question whether these data admit the

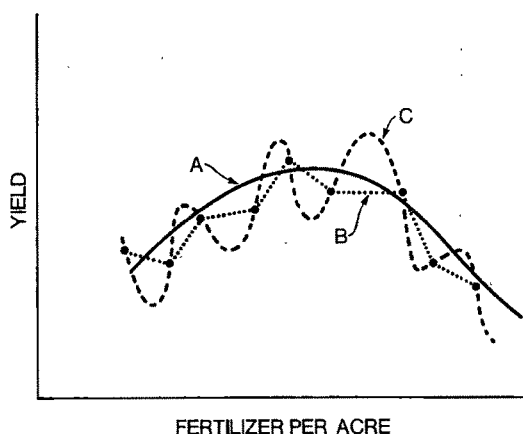


FIGURE 3. HYPOTHETICAL DATA AND THREE ESTIMATED FUNCTIONS

inference that the yield is maximized at  $F_m$ . Actually, after inspection of this figure, I don't think anything can be inferred except that the yield at  $F_2$  is higher than at  $F_1$ , which in turn is higher than at zero. Thus I don't believe the function is quadratic. If it is allowed to be a cubic then again there is an identification problem.

This kind of logic can be extended indefinitely. One can always find a set of observations that will make the inferences implied by a polynomial of degree  $p$  seem silly. This is true regardless of the degree  $p$ . Thus no model with a finite number of parameters is actually believed, whether the data are experimental or nonexperimental.

#### IV. Do We Need Prior Information?

A model with an infinite number of parameters will allow inference from a finite data set only if there is some prior information that effectively constrains the ranges of the parameters. Figure 3 depicts another hypothetical sequence of observations and three estimated relationships between yield and fertilizer. I believe the solid line *A* is a better representation of the relationship than either of the other two. The piecewise linear form *B* fits the data better, but I think this peculiar meandering function is highly unlikely on an a priori basis. Though *B* and *C* fit the data equally well, I believe that *B* is much more

likely than *C*. What I am revealing is the a priori opinion that the function is likely to be smooth and single peaked.

What should now be clear is that data alone cannot reveal the relationship between yield and fertilizer intensity. Data can reveal the yield at sampled values of fertilizer intensities, but in order to interpolate between these sampled values, we must resort to subjective prior information.

Economists have inherited from the physical sciences the myth that scientific inference is objective, and free of personal prejudice. This is utter nonsense. All knowledge is human belief; more accurately, human opinion. What often happens in the physical sciences is that there is a high degree of conformity of opinion. When this occurs, the opinion held by most is asserted to be an objective fact, and those who doubt it are labelled "nuts." But history is replete with examples of opinions losing majority status, with once-objective "truths" shrinking into the dark corners of social intercourse. To give a trivial example, coming now from California I am unsure whether fat ties or thin ties are aesthetically more pleasing.

The false idol of objectivity has done great damage to economic science. Theoretical econometricians have interpreted scientific objectivity to mean that an economist must identify exactly the variables in the model, the functional form, and the distribution of the errors. Given these assumptions, and given a data set, the econometric method produces an objective inference from a data set, unencumbered by the subjective opinions of the researcher.

This advice could be treated as ludicrous, except that it fills all the econometric textbooks. Fortunately, it is ignored by applied econometricians. The econometric art as it is practiced at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes. This searching for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference. The concepts of unbiasedness, consistency, efficiency, maximum-likelihood estimation,

in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose. The consuming public is hardly fooled by this chicanery. The econometrician's shabby art is humorously and disparagingly labelled "data mining," "fishing," "grubbing," "number crunching." A joke evokes the Inquisition: "If you torture the data long enough, Nature will confess" (Coase). Another suggests methodological fickleness: "Econometricians, like artists, tend to fall in love with their models" (wag unknown). Or how about: "There are two things you are better off not watching in the making: sausages and econometric estimates."

This is a sad and decidedly unscientific state of affairs we find ourselves in. Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. Like elaborately plumed birds who have long since lost the ability to procreate but not the desire, we preen and strut and display our *t*-values.

If we want to make progress, the first step we must take is to discard the counterproductive goal of objective inference. The dictionary defines an inference as a logical conclusion based on a set of facts. The "facts" used for statistical inference about  $\theta$  are first the data, symbolized by  $x$ , second a conditional probability density, known as a sampling distribution,  $f(x|\theta)$ , and, third, explicitly for a Bayesian and implicitly for "all others," a marginal or prior probability density function  $f(\theta)$ . Because both the sampling distribution and the prior distribution are actually *opinions* and not *facts*, a statistical inference is and must forever remain an *opinion*.

What is a fact? A fact is merely an opinion held by all, or at least held by a set of people you regard to be a close approximation to all.<sup>1</sup> For some that set includes only one

person. I myself have the opinion that Andrew Jackson was the sixteenth president of the United States. If many of my friends agree, I may take it to be a fact. Actually, I am most likely to regard it to be a fact if the authors of one or more books say it is so.

The difference between a fact and an opinion for purposes of decision making and inference is that when I use opinions, I get uncomfortable. I am not too uncomfortable with the opinion that error terms are normally distributed because most econometricians make use of that assumption. This observation has deluded me into thinking that the opinion that error terms are normal may be a fact, when I know deep inside that normal distributions are actually used only for convenience. In contrast, I am *quite* uncomfortable using a prior distribution, mostly I suspect because hardly anyone uses them. If convenient prior distributions were used as often as convenient sampling distributions, I suspect that I could be as easily deluded into thinking that prior distributions are facts as I have been into thinking that sampling distributions are facts.

To emphasize this hierarchy of statements, I display them in order: truths; facts; opinions; conventions. Note that I have added to the top of the order, the category truths. This will appeal to those of you who feel compelled to believe in such things. At the bottom are conventions. In practice, it may be difficult to distinguish a fact from a convention, but when facts are clearly unavailable, we must strongly resist the deceit or delusion that conventions can represent.

What troubles me about using opinions is their whimsical nature. Some mornings when I arise, I have the opinion that Raisin Bran is better than eggs. By the time I get to the kitchen, I may well decide on eggs, or oatmeal. I usually do recall that the sixteenth president distinguished himself. Sometimes I think he was Jackson; often I think he was Lincoln.

A data analysis is similar. Sometimes I take the error terms to be correlated, sometimes uncorrelated; sometimes normal and sometimes nonnormal; sometimes I include observations from the decade of the fifties, sometimes I exclude them; sometimes the

<sup>1</sup>This notion of "truth by consensus" is espoused by Thomas Kuhn (1962) and Michael Polanyi (1964). Oscar Wilde agrees by dissent: "A truth ceases to be true when more than one person believes it."

equation is linear and sometimes nonlinear; sometimes I control for variable  $z$ , sometimes I don't. Does it depend on what I had for breakfast?

As I see it, the fundamental problem facing econometrics is how adequately to control the whimsical character of inference, how sensibly to base inferences on opinions when facts are unavailable. At least a partial solution to this problem has already been formed by practicing econometricians. A common reporting style is to record the inferences implied by alternative sets of opinions. It is not unusual to find tables that show how an inference changes as variables are added to or deleted from the equation. This kind of sensitivity analysis reports special features of the mapping from the space of assumptions to the space of inferences. The defect of this style is that the coverage of assumptions is infinitesimal, in fact a zero volume set in the space of assumptions. What is needed instead is a more complete, but still economical way to report the mapping of assumptions into inferences. What I propose to do is to develop a correspondence between regions in the assumption space and regions in the inference space. I will report that all assumptions in a certain set lead to essentially the same inference. Or I will report that there are assumptions within the set under consideration that lead to radically different inferences. In the latter case, I will suspend inference and decision, or I will work harder to narrow the set of assumptions.

Thus what I am asserting is that the choice of a particular sampling distribution, or a particular prior distribution, is inherently whimsical. But statements such as "The sampling distribution is symmetric and unimodal" and "My prior is located at the origin" are not necessarily whimsical, and in certain circumstances do not make me uncomfortable.

To put this somewhat differently, an inference is not believable if it is fragile, if it can be reversed by minor changes in assumptions. As consumers of research, we correctly reserve judgment on an inference until it stands up to a study of fragility, usually by other researchers advocating opposite opinions. It is, however, much more efficient for

individual researchers to perform their own sensitivity analyses, and we ought to be demanding much more complete and more honest reporting of the fragility of claimed inferences.

The job of a researcher is then to report economically and informatively the mapping from assumptions into inferences. In a slogan, "The mapping is the message." The mapping does not depend on opinions (assumptions), but reporting the mapping economically and informatively does. A researcher has to decide which assumptions or which sets of alternative assumptions are worth reporting. A researcher is therefore forced either to anticipate the opinions of his consuming public, or to recommend his own opinions. It is actually a good idea to do both, and a serious defect of current practice is that it concentrates excessively on convincing one's self and, as a consequence, fails to convince the general professional audience.

The whimsical character of econometric inference has been partially controlled in the past by an incomplete sensitivity analysis. It has also been controlled by the use of conventions. The normal distribution is now so common that there is nothing at all whimsical in its use. In some areas of study, the list of variables is partially conventional, often based on whatever list the first researcher happened to select. Even conventional prior distributions have been proposed and are used with nonnegligible frequency. I am referring to Robert Shiller's (1973) smoothness prior for distributed lag analysis and to Arthur Hoerl and Robert Kennard's (1970) ridge regression prior. It used to aggravate me that these methods seem to find public favor whereas overt and complete Bayesian methods such as my own proposals (1972) for distributed lag priors are generally ignored. However, there is a very good reason for this: the attempt to form a prior distribution from scratch involves an untold number of partly arbitrary decisions. The public is rightfully resistant to the whimsical inferences which result, but at the same time is receptive to the use of priors in ways that control the whimsy. Though the use of conventions does control the whimsy, it can do so at the cost of relevance. Inferences based

on Hoerl and Kennard's conventional "ridge regression" prior are usually irrelevant, because it is rarely sensible to take the prior to be spherical and located at the origin, and because a closer approximation to prior belief can be suspected to lead to substantially different inferences. In contrast, the conventional assumption of normality at least uses a distribution which usually cannot be ruled out altogether. Still, we may properly demand a demonstration that the inferences are insensitive to this distributional assumption.

#### A. *The Horizon Problem: Sherlock Holmes Inference*

Conventions are not to be ruled out altogether, however. One can go mad trying to report completely the mapping from assumptions into inferences since the space of assumptions is infinite dimensional. A formal statistical analysis therefore has to be done within the limits of a reasonable horizon. An informed convention can usefully limit this horizon. If it turned out that sensible neighborhoods of distributions around the normal distribution 99 times out of 100 produced the same inference, then we could all agree that there are other more important things to worry about, and we may properly adopt the convention of normality. The consistency of least squares estimates under wide sets of assumptions is used improperly as support for this convention, since the inferences from a given finite sample may nonetheless be quite sensitive to the normality assumption.<sup>2</sup>

The truly sharp distinction between inference from experimental and inference from nonexperimental data is that experimental inference sensibly admits a conventional horizon in a critical dimension, namely the choice of explanatory variables. If fertilizer is randomly assigned to plots of land, it is conventional to restrict attention to the relationship between yield and fertilizer, and

to proceed as if the model were perfectly specified, which in my notation means that the misspecification matrix  $M$  is the zero matrix. There is only a small risk that when you present your findings, someone will object that fertilizer and light level are correlated, and there is an even smaller risk that the conventional zero value for  $M$  will lead to inappropriate inferences. In contrast, it would be foolhardy to adopt such a limited horizon with nonexperimental data. But if you decide to include light level in your horizon, then why not rainfall; and if rainfall, then why not temperature; and if temperature, then why not soil depth, and if soil depth, then why not the soil grade; ad infinitum. Though this list is never ending, it can be made so long that a nonexperimental researcher can feel as comfortable as an experimental researcher that the risk of having his findings upset by an extension of the horizon is very low. The exact point where the list is terminated must be whimsical, but the inferences can be expected not to be sensitive to the termination point if the horizon is wide enough.

Still, the horizon within which we all do our statistical analyses has to be ultimately troublesome, since there is no formal way to know what inferential monsters lurk beyond our immediate field of vision. "Diagnostic" tests with explicit alternative hypotheses such as the Durbin-Watson test for first-order autocorrelation do not truly ask if the horizon should be extended, since first-order autocorrelation is explicitly identified and clearly in our field of vision. Diagnostic tests such as goodness-of-fit tests, without explicit alternative hypotheses, are useless since, if the sample size is large enough, any maintained hypothesis will be rejected (for example, no observed distribution is exactly normal). Such tests therefore degenerate into elaborate rituals for measuring the effective sample size.

The only way I know to ask the question whether the horizon is wide enough is to study the anomalies of the data. In the words of the physiologist, C. Bernard:

A great surgeon performs operations for stones by a single method; later he

<sup>2</sup>In particular, least squares estimates are completely sensitive to the independence assumption, since by choice of sample covariance matrix a generalized least squares estimate can be made to assume any value whatsoever (see my 1981 paper).

makes a statistical summary of deaths and recoveries, and he concludes from these statistics that the mortality law for this operation is two out of five. Well, I say that this ratio means literally nothing scientifically, and gives no certainty in performing the next operation. What really should be done, instead of gathering facts empirically, is to study them more accurately, each in its special determinism...by statistics, we get a conjecture of greater or less probability about a given case, but never any certainty, never any absolute determinism...only basing itself on experimental determinism can medicine become a true science.

[1927, pp. 137-38]

A study of the anomalies of the data is what I have called "Sherlock Holmes" inference, since Holmes turns statistical inference on its head: "It is a capital mistake to theorize before you have all the evidence. It biases the judgements." Statistical theory counsels us to begin with an elicitation of opinions about the sampling process and its parameters; the theory, in other words. After that, data may be studied in a purely mechanical way. Holmes warns that this biases the judgements, meaning that a theory constructed before seeing the facts can be disastrously inappropriate and psychologically difficult to discard. But if theories are constructed after having studied the data, it is difficult to establish by how much, if at all, the data favor the data-instigated hypothesis. For example, suppose I think that a certain coefficient ought to be positive, and my reaction to the anomalous result of a negative estimate is to find another variable to include in the equation so that the estimate is positive. Have I found evidence that the coefficient is positive? It would seem that we should require evidence that is more convincing than the traditional standard. I have proposed a method for discounting such evidence (1974). Initially, when you regress yield on fertilizer as in equation (2), you are required to assess a prior distribution for the experimental bias parameter  $\beta^*$ ; that is, you must select the misspecification matrix  $M$ . Then, when the least squares estimate of  $\beta$

turns out to be negative, and you decide to include in the equation the light level as well as the fertilizer level, you are obligated to form a prior for the light coefficient  $\gamma$  consistent with the prior for  $\beta^*$ , given that  $\beta^* = \gamma r_1$ , where  $r_1$  is the regression coefficient of light on fertilizer.<sup>3</sup>

This method for discounting the output of exploratory data analysis requires a discipline that is lacking even in its author. It is consequently important that we reduce the risk of Holmesian discoveries by extending the horizon reasonably far. The degree of a polynomial or the order of a distributed lag need not be data instigated, since the horizon is easily extended to include high degrees and high orders. It is similarly wise to ask yourself before examining the data what you would do if the estimate of your favorite coefficient had the wrong sign. If that makes you think of a specific left-out variable, it is better to include it from the beginning.

Though it is wise to select a wide horizon to reduce the risk of Holmesian discoveries, it is mistaken then to analyze a data set as if the horizon were wide enough. Within the limits of a horizon, no revolutionary inference can be made, since all possible inferences are predicted in advance (admittedly, some with low probabilities). Within the horizon, inference and decision can be turned over completely to a computer. But the great human revolutionary discoveries are made when the horizon is extended for reasons that cannot be predicted in advance and cannot be computerized. If you wish to make such discoveries, you will have to poke at the horizon, and poke again.

## V. An Example

This rhetoric is understandably tiring. Methodology, like sex, is better demonstrated than discussed, though often better anticipated than experienced. Accordingly, let me give you an example of what all this

<sup>3</sup>In a randomized experiment with  $r_1 = 0$ , the constraint  $\beta^* = \gamma r_1$  is irrelevant, and you are free to play these exploratory games without penalty. This is a very critical difference between randomized experiments and nonrandomized nonexperiments.

ranting and raving is about. I trust you will find it even better in the experience than in the anticipation. A problem of considerable policy importance is whether or not to have capital punishment. If capital punishment had no deterrent value, most of us would prefer not to impose such an irreversible punishment, though, for a significant minority, the pure joy of vengeance is reason enough. The deterrent value of capital punishment is, of course, an empirical issue. The unresolved debate over its effectiveness began when evolution was judging the survival value of the vengeance gene. Nature was unable to make a decisive judgment. Possibly econometricians can.

In Table 1, you will find a list of variables that are hypothesized to influence the murder rate.<sup>4</sup> The data to be examined are state-by-state murder rates in 1950. The variables are divided into three sets. There are four deterrent variables that characterize the criminal justice system, or in economic parlance, the expected out-of-pocket cost of crime. There are four economic variables that measure the opportunity cost of crime. And there are four social/environmental variables that possibly condition the taste for crime. This leaves unmeasured only the expected rewards for criminal behavior, though these are possibly related to the economic and social variables and are otherwise assumed not to vary from state to state.

A simple regression of the murder rate on all these variables leads to the conclusion that each additional execution deters thirteen murders, with a standard error of seven. That seems like such a healthy rate of return, we might want just to randomly draft executees from the population at large. This proposal would be unlikely to withstand the scrutiny of any macroeconomists who are skilled at finding rational expectations equilibria.

The issue I would like to address instead is whether this conclusion is fragile or not. Does it hold up if the list of variables in the model is changed? Individuals with different experiences and different training will find

TABLE 1—VARIABLES USED IN THE ANALYSIS

- 
- |   |  |
|---|--|
| a. Dependent Variable                             | $M$ = Murder rate per 100,000, FBI estimate.   |
| b. Independent Deterrent Variables                | $PC$ = (Conditional) Probability of conviction for murder given commission. Defined by $PC = C/Q$ , where $C$ = convictions for murder, $Q = M \cdot NS$ , $NS$ = state population. This is to correct for the fact that $M$ is an estimate based on a sample from each state.<br>$PX$ = (Conditional) Probability of execution given conviction (average number of executions 1946–50 divided by $C$ ).<br>$T$ = Median time served in months for murder by prisoners released in 1951.<br>$XPOS$ = A dummy equal to 1 if $PX > 0$ .  |
| c. Independent Economic Variables                 | $W$ = Median income of families in 1949.<br>$X$ = Percent of families in 1949 with less than one-half $W$ .<br>$U$ = Unemployment rate.<br>$LF$ = Labor force participation rate.  |
| d. Independent Social and Environmental Variables | $NW$ = Percent nonwhite.<br>$AGE$ = Percent 15–24 years old.<br>$URB$ = Percent urban.<br>$MALE$ = Percent male.<br>$FAMHO$ = Percent of families that are husband and wife both present families.<br>$SOUTH$ = A dummy equal to 1 for southern states (Alabama, Arkansas, Delaware, Florida, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia).  |
| e. Weighting Variable                             | $SQRTNF$ = Square root of the population of the FBI-reporting region. Note that weighting is done by multiplying variables by $SQRTNF$ .   |
| f. Level of Observation                           | Observations are for 44 states, 35 executing and 9 nonexecuting. The executing states are: Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Illinois, Indiana, Kansas, Kentucky, Louisiana, Maryland, Massachusetts, Mississippi, Missouri, Nebraska, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, South Dakota, Tennessee, Texas, Virginia, Washington, West Virginia.<br>The nonexecuting states are: Idaho, Maine, Minnesota, Montana, New Hampshire, Rhode Island, Utah, Wisconsin, Wyoming. |
- 

<sup>4</sup>This material is taken from a study by a student of mine, Walter McManus (1982).

different subsets of the variables to be candidates for omission from the equation. Five different lists of doubtful variables are reported in Table 2. A right winger expects

TABLE 2—ALTERNATIVE PRIOR SPECIFICATIONS

Prior	PC	PX	T	XPOS	W	X	U	LF	NW	AGE	URB	MALE	FAMHO	SOUTH
Right Winger	I	I	I	*	D	D	D	D	D	D	D	D	D	D
Rational Maximizer	I	I	I	*	I	I	I	I	D	D	D	D	D	D
Eye-for-an-Eye	I	I	D	*	D	D	D	D	D	D	D	D	D	D
Bleeding Heart	D	D	D	*	I	I	I	I	D	D	D	D	D	D
Crime of Passion	D	D	D	*	I	I	I	I	I	I	I	I	I	I

Notes: 1) *I* indicates variables considered important by a researcher with the respective prior. Thus, every model considered by the researcher will include these variables. *D* indicates variables considered doubtful by the researcher. \* indicates *XPOS*, the dummy equal to 1 for executing states. Each prior was pooled with the data two ways: one with *XPOS* treated as important, and one with it as doubtful.

2) With five basic priors and *XPOS* treated as doubtful or important by each, we get ten alternative prior specifications.

the punishment variables to have an effect, but treats all other variables as doubtful. He wants to know whether the data still favor the large deterrent effect, if he omits some of these doubtful variables. The rational maximizer takes the variables that measure the expected economic return of crime as important, but treats the taste variables as doubtful. The eye-for-an-eye prior treats all variables as doubtful except the probability of execution. An individual with the bleeding heart prior sees murder as the result of economic impoverishment. Finally, if murder is thought to be a crime of passion then the punishment variables are doubtful.

In Table 3, I have listed the extreme estimates that could be found by each of these groups of researchers. The right-winger minimum of -22.56 means that a regression of the murder rate data on the three punishment variables and a suitably selected linear combination of the other variables yields an estimate of the deterrent effect equal to 22.56 lives per execution. It is possible also to find an estimate of -.86. Anything between these two extremes can be similarly obtained; but no estimate outside this interval can be generated no matter how the doubtful variables are manipulated (linearly). Thus the right winger can report that the inference from this data set that executions deter murders is not fragile. The rational maximizer similarly finds that conclusion insensitive to choice of model, but the other three priors allow execution actually to encourage murder, possibly by a brutalizing effect on society.

TABLE 3—EXTREME ESTIMATES OF THE EFFECT OF EXECUTIONS ON MURDERS

Prior	Minimum Estimate	Maximum Estimate
Right Winger	-22.56	-.86
Rational Maximizer	-15.91	-10.24
Eye-for-an-Eye	-28.66	1.91
Bleeding Heart	-25.59	12.37
Crime of Passion	-17.32	4.10

Note: Least squares is -13.22 with a standard error of 7.2.

I come away from a study of Table 3 with the feeling that any inference from these data about the deterrent effect of capital punishment is too fragile to be believed. It is possible credibly to narrow the set of assumptions, but I do not think that a credibly large set of alternative assumptions will lead to a sharp set of estimates. In another paper (1982), I found a narrower set of priors still leads to inconclusive inferences. And I have ignored the important simultaneity issue (the death penalty may have been imposed in crime ridden states to deter murder) which is often a source of great inferential fragility.

## VI. Conclusions

After three decades of churning out estimates, the econometrics club finds itself under critical scrutiny and faces incredulity as never before. Fischer Black writes of "The Trouble with Econometric Models." David

Hendry queries "Econometrics: Alchemy or Science?" John W. Pratt and Robert Schlaifer question our understanding of "The Nature and Discovery of Structure." And Christopher Sims suggests blending "Macroeconomics and Reality."

It is apparent that I too am troubled by the fumes which leak from our computing centers. I believe serious attention to two words would sweeten the atmosphere of econometric discourse. These are whimsy and fragility. In order to draw inferences from data as described by econometric texts, it is necessary to make whimsical assumptions. The professional audience consequently and properly withholds belief until an inference is shown to be adequately insensitive to the choice of assumptions. The haphazard way we individually and collectively study the fragility of inferences leaves most of us unconvinced that any inference is believable. If we are to make effective use of our scarce data resource, it is therefore important that we study fragility in a much more systematic way. If it turns out that almost all inferences from economic data are fragile, I suppose we shall have to revert to our old methods lest we lose our customers in government, business, and on the boardwalk at Atlantic City.

## REFERENCES

- Bernard, C., *An Introduction to the Study of Experimental Method*, New York: MacMillan, 1927.
- Black, Fischer, "The Trouble with Econometric Models," *Financial Analysts Journal*, March/April 1982, 35, 3-11.
- Friedman, Milton, *A Theory of the Consumption Function*, Princeton: Princeton University Press, 1957.
- Hendry, David, "Econometrics—Alchemy or Science?," *Economica*, November 1980, 47, 387-406.
- Hoerl, Arthur E. and Kennard, Robert W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, February 1970, 12, 55-67.
- Kuhn, Thomas S., *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press, 1962.
- Lakatos, Imre, "Falsification and the Methodology of Scientific Research Programmes," in his and A. Musgrave, eds., *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 1969.
- Leamer, Edward E., "A Class of Prior Distributions and Distributed Lag Analysis," *Econometrica*, November 1972, 40, 1059-81.
- , "False Models and Post-data Model Construction," *Journal American Statistical Association*, March 1974, 69, 122-31.
- , *Specification Searches: Ad Hoc Inference with Non-experimental Data*, New York: Wiley, 1978.
- , "Techniques for Estimation with Incomplete Assumptions," *IEEE Conference on Decision and Control*, San Diego, December 1981.
- , "Sets of Posterior Means with Bounded Variance Priors," *Econometrica*, May 1982, 50, 725-36.
- McManus, Walter, "Bayesian Estimation of the Deterrent Effect of Capital Punishment," mimeo., University of California-Los Angeles, 1981.
- Plott, Charles R. and Smith, Vernon L., "An Experimental Examination of Two Exchange Institutions," *Review of Economic Studies*, February 1978, 45, 133-53.
- Polanyi, Michael, *Personal Knowledge*, New York: Harper and Row, 1964.
- Pratt, John W. and Schlaifer, Robert, "On the Nature and Discovery of Structure," mimeo., 1979.
- Shiller, Robert, "A Distributed Lag Estimator Derived From Smoothness Priors," *Econometrica*, July 1973, 41, 775-88.
- Sims, C. A., "Macroeconomics and Reality," *Econometrica*, January 1980, 48, 1-48.
- , "Scientific Standards in Econometric Modeling," mimeo., 1982.
- Smith, Vernon L., "Relevance of Laboratory Experiments to Testing Resource Allocation Theory," in J. Kmenta and J. Ramsey, eds., *Evaluation of Econometric Models*, New York: Academic Press, 1980, 345-77.

# Why Real Interest Rates Were So Low in the 1970's

By JAMES A. WILCOX\*

The economic turmoil of the 1970's resulted in record postwar increases in inflation, unemployment, and nominal interest rates. Yet, it was declines that were hardest to explain. The average real value of a share of common stock plummeted. Productivity growth evaporated. Real interest rates, measured as the spread between nominal interest rates and the inflation rate, turned negative (see Table 1).

Over the past decade, an explosion has occurred in the amount of attention paid to the relationship between nominal and thus real interest rates and expected inflation rates. A good deal of this effort has been directed toward empirical analysis of the Fisher neutrality hypothesis that nominal rates respond one for one with expected inflation rates. Most empirical tests of the Fisher hypothesis have been bivariate; interest rates were regressed on a constant and on actual or expected inflation measures. Estimates of the impact of inflation on interest rates in these and even in extended models are often significantly below one, are often statistically imprecise, and tend to be unstable over time.<sup>1</sup> The middle column of Table 1 reflects this. The coefficient of inflation on nominal interest rates there drops from 0.78 to 0.59 in the latter 1970's.

Another branch of work on nominal interest rates has concentrated on the institutional impediments to the Fisher hypothesis. Robert Mundell and James Tobin demon-

strate that nominal rates change by a smaller amount than the expected inflation rate does when a real balance effect exists and money pays no interest. Michael Darby and Martin Feldstein, on the other hand, argue that nominal rates should exhibit a greater-than-unity response to expected inflation due to the nature of U.S. income tax laws.<sup>2</sup>

Here I estimate the relation between interest rates, expected inflation, and real forces. Such estimates are required for evaluating the effects of inflation on saving, investment, the distribution of income, and the redistribution of wealth. To generate estimates of the net impact of these various factors on interest rates and to avoid the identification and simultaneity problems masked, but not often remedied, by instrumental variables techniques, I employ only exogenous regressors. To buttress the argument, I examine the individual links in the chain which are summarized in the reduced form.

The novel aspect of the model presented in Section I is its addition of aggregate supply shocks to the determination of interest rates. I trace the reduction in the supply of complementary factor inputs in the 1970's to a decline in the demand for capital and therefore in real interest rates. The inclusion of this supply force along with expected inflation allows one to distinguish between two, offsetting effects on interest rates: the depressing effect on real rates through lower investment demand and the elevating effect on nominal rates of higher expected inflation.

The results in Section II not only strongly support the economic and statistical importance of supply shocks on real interest rates, but also resolve some longstanding interest rate puzzles. Allowing for the impact of factor supply produces a significant estimated response to expected inflation for a sample that ends prior to late 1960's. The magnitude

\*Assistant professor, School of Business Administration, University of California-Berkeley. I would like to thank George Akerlof, Robert J. Gordon, Robert A. Meyer, Joe Peek, Janet Yellen, the members of the Economic Analysis and Policy seminar at Berkeley, participants at the NBER Conference on Inflation and Financial Markets, and anonymous referees for helpful comments. Data Resources supplied data and computational services. Financial support was provided by the Berkeley Program in Finance and NSF grant SES-8109093. Linda Pacheco supplied able research assistance. Errors of omission or commission are my responsibility.

<sup>1</sup>See William Gibson, David Pyle, John Carlson, and Thomas Cargill and Robert Meyer.

<sup>2</sup>Maurice Levi and John Makin derive the reduced-form effect of expected inflation on nominal interest rates as a function of the structural parameters.

TABLE 1—HISTORICAL AVERAGES

	Nominal Interest Rate	Inflation Rate	Real Interest Rate	Response of Nominal Interest to Inflation Rate	Relative Price of Energy	Pre-Tax Profit Rate	Capital Stock Growth Rate
1960-73	4.6	3.3	1.3	0.78	1.00	12.9	4.2
1974-78	6.5	7.3	-0.8	0.59	1.59	9.4	2.7

*Notes; Sources:* The nominal interest rate is the annual average rate on six-month Treasury bills (Table B-67). The inflation rate is the percentage change in the annual GNP deflator (Table B-5) from the previous year. The real interest rate is their difference. The responses of nominal interest rates to inflation come from a 1960-78 annual data OLS regression of nominal rates on a constant and two variables. The 1960-73 response is the coefficient on a variable which takes the values of the inflation rate for 1960-73 and is zero otherwise. The 1974-78 response is the coefficient on a variable that takes the values of inflation for 1974-78 and is zero otherwise. The relative price of energy is an index (1960-73 = 1.00) of the ratio of the producer price of fuel and power (Table B-57) to the GNP deflator. All of the above series are from the 1982 *Economic Report of the President*. Pre-tax profit rates are from Daniel Holland and Stewart Myers (Table 1). The capital stock growth rate is from Feldstein (p. 20).

of that response is very similar to that estimated for a later period. Failure to untangle the offsetting effects of supply forces on nominal rates has also led in the past to unwarranted claims of coefficient instability. Incorporation of supply factors, by contrast, results in a single, stable interest rate function for the entire post-Accord period, as demonstrated in Section IV. In particular, the apparently declining effect of expected inflation on interest rates, especially in the 1970's, is shown to result from failure to allow for changing factor supply forces.

### I. A Model of Interest Rates

To explain the movement of interest rates over time, I posit a simple macro model. Commodity market equilibrium is given by an *IS* curve in inverse form:<sup>3</sup>

$$(1) \quad r_{at}^e = a_0 + a_1 X - a_2 Q + a_3 (M - P) - a_4 SS.$$

<sup>3</sup>The following variable definitions are used in (1)-(5):  $r_{at}^e$  = real, after-tax interest rate;  $i$  = nominal interest rate;  $t$  = constant marginal tax rate;  $X$  =  $\log$  of exogenous real demand, normalized by natural real output;  $Q$  =  $\log$  of real output, normalized by natural real output;  $M$  =  $\log$  of exogenous nominal money supply, normalized by natural real output;  $P$  =  $\log$  of general price level,  $P^e$  =  $\log$  of expected general price level;  $p^e$  = expected inflation rate; and  $SS$  = supply shock proxy. All parameters are assumed to be positive.

The relevant interest rate for household and firm expenditure decisions is the real after-tax rate, defined as

$$(2) \quad r_{at}^e = i(1 - t) - p^e.$$

The *IS* curve shifts in response to changes in exogenously determined spending, proxied by  $X$ . The Mundell-Tobin effect is allowed for by including a real balance effect.

A reduction in the supply of one input, like energy, reduces use of that factor and reduces the marginal products of the remaining inputs. This coincides with the fall of pre-tax profit rate in the later 1970's shown in Table 1. Even in a Cobb-Douglas setting, where the remaining inputs (for example, capital and labor) tend to be substituted for the dearer input, the gross substitution effect is more than offset by the (negative) expansion effect.<sup>4</sup> Several studies have estimated aggregate production functions with functional forms which allow for varying degrees of substitutability among factor inputs. Ernst

<sup>4</sup>See my earlier paper for a detailed example. John Tatom also shows that the net effect is a decline in the demand for capital. Leif Johansen proves a similar result in a putty-clay framework. Supply shocks may induce a shift in the supply of capital as well. Two possible, offsetting effects are a reduction due to a decline in equilibrium value-added and an increase in the propensity to save associated with the redistribution of wealth.

Berndt and David Wood, Berndt and Mohammed Khaled, and Edward Hudson and Dale Jorgenson have each found the energy and capital to be net complements empirically. These results imply that the net demand for capital falls in response to such supply shifts.<sup>5</sup>

Both the growth rate of investment and its share of total output did fall dramatically all over the industrialized world starting almost exactly at the time real energy prices rose.<sup>6</sup> Investment also remained weak throughout the recovery from the mid-1970's recession. The 1978 *Economic Report of the President* notes (p. 66) that whereas real business fixed investment (*BFI*) had increased an average of 14 percent above its level at the previous cyclical peak at a similar stage in other post-war recoveries, in late 1977, real *BFI* was still 2 percent below its level at the 1973 cycle peak. Table 1 documents the appreciably lower capital stock growth rate that such a negative supply shock would imply.

To the extent that supply shocks are responsible for the lower level of investment, empirical models of investment that are driven primarily by the cost of capital should exhibit a tendency to overpredict after negative supply shocks strike. The neoclassical models estimated by Richard Kopcke and by Peter Clark, that do not allow for factor supply changes, do precisely that after 1974.<sup>7</sup> Kopcke further concludes that "all models seriously overpredict real *BFI*..." (p. 25).

Most models embody slides along the capital, and thus investment, demand schedules associated with interest rate changes. A supply shock, like cartelization of the world petroleum supply, reduces the demand for capital and, thereby, interest rates. In addition to the slides along the schedules, my

model captures shifts in those demand schedules occasioned by changes in factor supplies with a supply shock proxy, *SS*. It is these shifts that prior empirical models of investment and interest rates have missed.

Equilibrium in the money market is characterized by an *LM* curve in inverse form:

$$(3) \quad Q = b_0 + b_1(M - P) + b_2i(1 - t).$$

The model is closed with the addition of an aggregate supply (*AS*) function:

$$(4) \quad P = P^e + cQ.$$

The reduced-form implies nominal interest rates depend upon expected inflation, on exogenous demand, on liquidity forces, and on aggregate supply conditions:

$$(5) \quad i = 1/D(1 - t)[a_0(1 + cb_1) - b_0(a_2 + ca_3) + (1 + cb_1)P^e + (a_3 - a_2b_1)(M - P^e) + a_1(1 + cb_1)X - a_4(1 + cb_1)SS];$$

where  $D \equiv 1 + cb_1 + a_2b_2 + cb_2a_3$ .

The model predicts that the coefficient on exogenous spending, *X*, will be positive. The response of nominal interest rates to expected inflation will also be positive, but not necessarily greater or less than one. Higher effective marginal tax rates drive up that coefficient, but the *IS*, *LM*, and *AS* slopes tend to produce a coefficient that is less than one. Thus we may observe responses that are either greater or less than unity, depending on the size of various parameters. The sign of the coefficient on the exogenous liquidity variable is ambiguous; the Mundell-Tobin effect counters the depressing effect on real rates of more liquidity. A priori, we cannot know which effect dominates. I anticipate the supply shock variable, *SS*, will carry a negative coefficient.

To date, most analyses of aggregate supply shocks have concentrated on short-run effects, with particular emphasis on inflation.<sup>8</sup>

<sup>5</sup>Hudson and Jorgenson estimate a reduction in the desired capital stock from 1972 to 1976 of about \$100 billion. Increases in relative energy prices after 1976 would imply an even larger reduction.

<sup>6</sup>See Jeffrey Sachs and Otto Eckstein. The relative price of energy averaged 60 percent higher after 1973 (see Table 1).

<sup>7</sup>The neoclassical models also forecast the (out-of-sample) post-1973 era poorly relative to other models, in spite of their tendency to outperform other models in-sample.

<sup>8</sup>See Gordon (1975) and Edmund Phelps.

Given an imperfectly flexible nominal wage, an increase in the price of materials raises the aggregate supply schedule in price-output space. *Ceteris paribus*, this lowers the real money supply and thereby raises the real interest rate and lowers investment and output. We grant that short-run frictions may lead to temporarily higher real interest rates and lower investment.<sup>9</sup> The thrust of my model, however, is that permanent real supply shocks reduce the return to and demand for capital (and to labor) for the longer run. The work of Robert J. Gordon (1980) and Edward Gramlich can be interpreted as demonstrating this equilibrium effect on real wages empirically for the United States during the 1970's.

## II. Empirical Results

To test the hypothesis that negative real supply shocks drive down real and, *ceteris paribus*, nominal interest rates, I estimate a reduced-form equation based on (5). My sample extends from 1952 through 1979. The interval begins with the Federal Reserve's cessation of interest rate pegging and ends just prior to the imposition of credit controls in 1980.<sup>10</sup> The interest rate,  $i$ , is the annual nominal yield on one-year Treasury bills.<sup>11</sup> I

<sup>9</sup>We can conceive of movements of aggregate output away from its equilibrium level as being generated by movements of the actual labor supply function from its equilibrium level. Recessions then result from upward shifts of the labor supply function relative to the real wage. This withdrawal of labor corresponds to the upward shift of the materials supply function and drives down the return to capital in an analogous fashion. Thus, we expect the real rate of interest to fluctuate with the business cycle. Labor supply could, of course, exogenously shift. The shifts in labor supply that generate cyclical output patterns result from misperceptions, however, which are assumed to be eliminated in the long run.

<sup>10</sup>The structural changes engendered by the enactment of the Depository Institutions Deregulation and Monetary Control Act of 1980 and the imposition of credit controls in early 1980 preclude extending the sample past 1979. Restricting the sample to the post-1953 period to allow the effects of interest rate pegging to dissipate did not change my results perceptibly.

<sup>11</sup>Before December 1959, when one-year bills were introduced, the rate is the yield on bills with maturities of nine to twelve months. Since the inflation forecasts were made by early June and early December, I use

use Treasury bill rates, as opposed to commercial paper rates, for example, in order to minimize coefficient biases and maximize efficiency.<sup>12</sup> The differential between these two series is far from constant. To the extent that the changes in the differential, presumably due to market perceptions of relative default risk and liquidity, are correlated with any of the explanatory variables, estimates based on commercial paper rates will be biased.<sup>13</sup> Even if the differential is orthogonal so that the point estimates are unbiased, the larger the variance of the differential, the less efficient the coefficient estimates will be.

The independent expected inflation series,  $PE12$ , is the Livingston series and is available semiannually for a twelve-month horizon.<sup>14</sup> Among the advantages of using the Livingston data are that, 1) they are not contaminated by future information as full sample, least squares estimates of expectations are; 2) they may incorporate highly nonlinear and varying parameter schemes; and 3) they are based upon broader information sets than regression-based estimates of expected inflation.

Shifts in the  $LM$  curve are captured by a proxy variable for liquidity,  $LIQ$ , which is defined as the annualized growth rate of the nominal money supply over the last six months minus its annualized growth rate over the last three years.<sup>15</sup> This specification allows for the tendency of financial asset markets to clear more rapidly than the physi-

June and December monthly averages for interest rates. Second- and fourth-quarter data are used for all other series.

<sup>12</sup>Carlson uses commercial paper rates. Cargill uses Treasury bill rates.

<sup>13</sup>The difference between the four-to-six month commercial paper and the six-month Treasury bill rates over the 1959-78 portion of my sample is significantly related to the expected inflation rate for the next six months, and to the liquidity variable in a regression that also includes the supply proxy and exogenous demand.

<sup>14</sup>The Federal Reserve Bank of Philadelphia generously provided this series.

<sup>15</sup>This is the same specification used by Carlson. Cargill and Meyer use a similar measure, but do not detrend money growth. At various stages, the growth rate of real balances and the inverse of expected natural velocity were substituted for detrended money growth. In general, the detrended money measure,  $LIQ$ , was associated with a slightly better overall fit.

cal output market. An increase in money growth initially depresses real rates due to the dominance of the liquidity effect.<sup>16</sup> The ensuing rise in income raises rates while *LIQ* recedes as the higher money growth rate is maintained. Eventually output and *LIQ* each revert to their original equilibrium levels as prices and expected inflation adapt to the higher long-run money growth rate. My use of monthly average interest rate data makes this dynamic response due to differential adjustment speeds very likely.

I capture movements of the *IS* curve with two variables besides the liquidity factor. The first, *X*, is designed to pick up autonomous shifts in demand which are unrelated to supply shocks. Here, *X* is the sum of real federal government defense expenditures and real exports, normalized by natural real output.<sup>17</sup> The second variable that contributes to *IS* curve shifts, *SUPPLY*, proxies changes in the world supply of materials, broadly defined to include both materials for further processing like raw copper, and inputs consumed in production, like oil, and is measured by the ratio of the implicit price deflator for imports to the *GNP* deflator.<sup>18</sup> The import deflator responds either to supply shifts or changes in the exchange rate. Exchange rate changes and real materials supply shocks may have differential effects on real interest rates, since the former implies an offsetting shock to foreign markets while the latter may be a positive or negative shock to all. I strip exchange rate shocks from the import deflator by multiplying it by the effective exchange rate.<sup>19</sup>

Table 2 contains the results of estimating truncated and complete versions of the reduced form for interest rates:

$$(6) \quad i = \beta_0 + \beta_1 PE12 + \beta_2 LIQ \\ + \beta_3 X + \beta_4 SUPPLY.$$

In each case the full sample is employed and Cochrane-Orcutt autocorrelation-corrected (*CORC*) estimates are presented. The first row gives the results for a skeletal Fisher equation; only expected inflation appears as an explanatory variable. The estimated coefficient on expected inflation of 0.76 is significantly below 1.00. Though the specification in row 1 has been widely used to test the Fisher neutrality hypothesis and a less-than-unity coefficient has been reported many times, equation (5) points out that it is seriously incomplete. The resulting omitted variable bias of unknown direction compromises any conclusion based on a skeletal specification. Recently, expanded versions have been estimated, similar to row 2 where the liquidity variable *LIQ* and exogenous demand *X* have been added. The effect of *LIQ* is insignificantly negative. Since the model points to offsetting forces on rates from liquidity pressures, this insignificance is not surprising.<sup>20</sup> Exogenous demand has a positive effect on real rates at a significance level of 6 percent. As in row 1, a large residual autocorrelation correction ( $\rho = 0.63$ ) is required in row 2. Though the true errors may be autocorrelated, our model suggests that mis-

<sup>16</sup>See Darby and Milton Friedman.

<sup>17</sup>Natural real output is based on Robert Rasche and John Tatom, as updated by Tatom (personal communication).

<sup>18</sup>The variables *SUPPLY* and *X* are measured as deviations from their respective 1952-79 sample means.

<sup>19</sup>This allows us to enter the effective exchange rate separately and determine if offsetting exchange rate shocks affect domestic interest rates. The estimated effect turns out to be insignificant. The lack of a significant impact of exchange rate changes is consistent with integration of international capital markets. Domestic rates may be virtually immune to adverse relative shifts in input prices occasioned by exchange rate changes if capital flows toward the country experiencing the corresponding favorable shift in exchange rates. This upward

shift in domestic capital supply will tend to offset the depressing effect on interest rates that operates through the demand for capital.

<sup>20</sup>When we substitute endogenous total real demand (real output relative to natural real output) for exogenous demand *X*, we find that its coefficient is highly significant ( $t = 5.01$ ) and that the expected inflation coefficient is virtually unchanged (0.811). The significance of total demand presumably reflects the simultaneity bias occasioned by commodity market shocks and the lagged, endogenous response of interest rates and output to *LM* curve shifts. To the extent the former is true, exogenous demand is the preferred variable. To the extent the latter is true, the specification of *LIQ* is supported. Nearly all prior work uses endogenous total output rather than exogenous demand, which leads to biases of unknown size.

TABLE 2—EFFECTS OF EXPECTED INFLATION, LIQUIDITY, EXOGENOUS DEMAND, AND SUPPLY SHOCKS ON NOMINAL INTEREST RATES; SEMIANNUALLY  
(*t*-statistics in parentheses)

Sample Period	Constant	PE12	LIQ	X	SUPPLY	RPENERGY	$\rho$	R <sup>2</sup>	D.W.	S.E.E.
1. 1952:06–1979:12	2.59 (7.07)	0.761 (8.74)	—	—	—	—	0.57 (5.20)	0.8686	1.88	0.805
2. 1952:06–1979:12	2.43 (5.83)	0.845 (8.17)	–3.74 (–0.64)	31.4 (1.94)	—	—	0.63 (6.03)	.8794	2.10	0.786
3. 1952:06–1979:12	1.96 (7.63)	0.987 (12.85)	–4.70 (–0.84)	43.7 (3.71)	–5.99 (–4.59)	—	0.36 (2.83)	.9070	2.02	0.697
4. 1952:06–1979:12	1.74 (4.69)	1.081 (9.02)	–3.82 (–0.64)	35.1 (2.63)	—	–1.78 (–2.72)	0.43 (3.51)	.8895	2.02	0.761
5. 1952:06–1965:12	2.13 (4.57)	1.066 (2.50)	–11.41 (–1.18)	39.3 (1.61)	–8.91 (–2.13)	—	0.31 (1.72)	.6203	1.93	0.698
6. 1966:06–1979:12	1.48 (2.44)	1.113 (9.48)	–5.76 (–0.87)	67.0 (5.60)	–6.01 (–4.24)	—	–0.01 (–0.07)	.8597	2.05	0.656

*Sources:* The originating agency is listed for series available on the *DRI* data base system. The nominal money supply is *M1A* and interest rates are the average market yields on one-year bills (Board of Governors of the Federal Reserve System). Federal government defense purchases and its deflator, exports of goods and services and its deflator, the deflator for imports, and the *GNP* deflator are from the U.S. Department of Commerce. The producer price index for fuel and power is from the U.S. Department of Labor. The trade-weighted exchange rate index is calculated by Morgan Guarantee Trust. See text for variable definitions.

specification in the form of an omitted variable has induced a portion of that autocorrelation.

Row 3 adds the shock proxy. The coefficient on *SUPPLY* is significant ( $t = -4.59$ ) and, as hypothesized, negative: supply shocks which raise the relative price of materials drive down real rates even after allowing for liquidity and exogenous demand effects. Further, adding *SUPPLY* raises the significance of the exogenous demand effect, reduces the extent of residual autocorrelation, and lowers the standard error of the estimate and of each coefficient. Including *SUPPLY* also raises the expected inflation coefficient, from 0.845 to 0.987. Thus, omitting *SUPPLY* leads to a downward bias in the estimated response of nominal rates to expected inflation.

Other forms of the proxy variable for supply shocks were considered as well. The petroleum and agricultural sectors are the sites of the most dramatic recent supply shocks. To capture the exogenous shocks emanating from these sectors specifically, the relative domestic prices of energy and of food were each entered separately and at the same time. In general, the relative price of food proved insignificant and the relative price of energy significant, but the overall fits were better with the broader relative

price proxy of row 3. Row 4 substitutes *RPENERGY*, the ratio of the producer price index for fuel and power to the *GNP* deflator, for *SUPPLY*.<sup>21</sup> The coefficient on *RPENERGY* is significant and negative and, again, the expected inflation coefficient rises appreciably relative to that in row 2. To the extent that supply shocks, droughts, or increases in the exercise of monopoly power occur in other sectors, however, single-sector-based proxies are too narrow. Because of this concern that broader measures might be more accurate proxies, I retain the relative price of imports adjusted for exchange rate changes as my preferred measure.

### III. Real Rate Movements

The thrust of (5) is that a bivariate specification is incomplete, since interest rates vary systematically with other forces. Eugene Fama (1975, 1977) argues that, nonetheless, the preponderance of market interest rate movement is due to changes in expected inflation. The *t*-statistics of Table 2 support this conclusion in that the largest *t*-values are always associated with expected inflation. Further, the simple correlation between the

<sup>21</sup>The correlation between *SUPPLY* and *RPENERGY* is 0.71.

nominal rate and the predicted component due to expected inflation is 0.90 while the correlation between the nominal rate and the predicted component due to all the remaining factors is  $-0.39$ .<sup>22</sup>

This is not to say that changes in real rates have been inconsequential. Frederic Mishkin (1981) argues that real rates have varied considerably over time, but he is unable to find any significant relation between real rates and either real or monetary forces once expected inflation is taken into account. In that circumstance, one can readily agree with him that "there is something left to explain."<sup>23</sup> Table 2 indicates that fiscal policy and aggregate supply forces are two factors that do help explain real rates.

Using (2), the results in row 3 imply<sup>24</sup>

$$(7) \quad r_{at}^e = \begin{matrix} 1.33 & -0.329 & & -3.20 \\ (7.63) & (-6.30) & & (-0.84) \end{matrix} \begin{matrix} PE12 \\ LIQ \end{matrix} \\ + \begin{matrix} 29.7 & X & -4.07 \\ (3.71) & & (-4.59) \end{matrix} \begin{matrix} \\ SUPPLY \end{matrix}.$$

The "steady-state" expected after-tax real rate  $r_{at}^e$  of 1.33 percent is significantly positive at well over 99 percent confidence.<sup>25</sup> This steady-state rate is defined as that interest rate predicted by my estimates when expected inflation is zero, money growth proceeds at its recent trend, exogenous spending as a fraction of potential real GNP is at its sample average, and the supply shock variable is likewise at its sample average. Since we may observe nonzero expected inflation forever, and since the components of  $X$  and  $SUPPLY$  need not ever return to the sample average values, there is no presumption that

this rate will ever reign or even be approached.

Changes in rates occasioned by these forces remain correctly measured, however, regardless of the base used for steady-state calculations. Figure 1 plots the effect on expected real after-tax interest rates of changing expected inflation rates ( $PE12$ ) and supply conditions ( $SUPPLY$ ) and allows us to assess the economic significance of each. Each plotted series traces the impact of a variable on  $r_{at}^e$  and is calculated as the product of the variable and its coefficient from row 3, less the value of that product for the first observation of the sample. Thus, each measure's movement of the expected real rate is due to that factor from the 1952:6 base.

Figure 1 shows that the declining relative price of materials produced a continuing positive supply shock from 1952 until the early 1970's. In part this reflects the decline of the relative price of oil over this period. Saudi Arabian crude oil, for example, sold for \$1.71 in 1950, but only \$1.30 in 1970.<sup>26</sup> Over that same period, the U.S. aggregate price level went up 70 percent. The result was a 55 percent decline in the relative price of Saudi crude. This positive supply shock raised the after-tax real rate by over one and one-half percentage points from the early 1950's until the late 1960's. The steadily rising expected inflation rate, by contrast, drove  $r_{at}^e$  down by a full percentage point by the end of the 1960's.

Until the early 1970's, then, the positive supply effect more than offset the depressing effect of higher expected inflation. The advent of negative supply shocks starting in the early 1970's, however, reinforced the downward pressure on real rates. Through the 1970's, supply forces drove down real rates by over one percentage point. At the same time, the rising expected inflation rate depressed the real rate of interest another one and one-half percentage points. Prior to that time, the net effect of these two forces had been to raise real returns by a fairly small amount. This may account for the lack of attention paid to these important factors

<sup>22</sup>These correlations are based on the coefficients in row 3.

<sup>23</sup>The variance of the expected real rate ( $i - PE12$ ) over the 1952-79 sample is 1.3 percentage points. The mean is 1.9.

<sup>24</sup>I use a constant average (across incomes) marginal interest-income tax rate of 32 percent. This is the average for the 1952-75 period calculated by Vito Tanzi (1980). The coefficient on  $PE12$  in (7) is very sensitive to the tax rate assumed. A zero effective rate implies of a coefficient of  $(0.987 - 1.000) = -0.013$ .

<sup>25</sup>Given the construction of my variables, the steady-state rate is estimated by the regression constant term.

<sup>26</sup>International Financial Statistics, 1980 Yearbook.

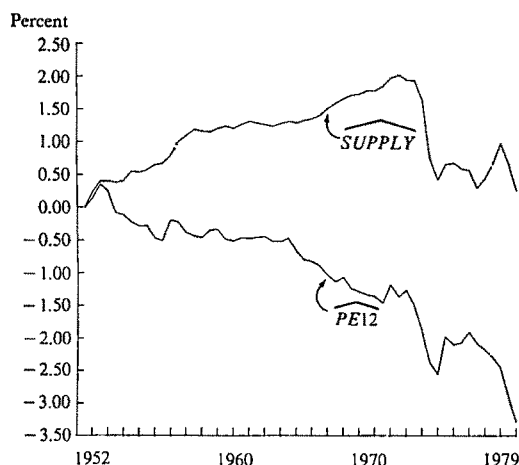


FIGURE 1. THE MOVEMENT OF EXPECTED REAL AFTER-TAX INTEREST RATES DUE TO CHANGING EXPECTED INFLATION RATES AND SUPPLY CONDITIONS 1952:6–1979:12, SEMIANNUALLY (1952:6 = 0.0)

during this time. By the late 1970's, however, these two factors were each driving down the real rate. The result was a sustained period of historically low real rates which existing models could not explain.

#### IV. Stability of the Interest Rate Relation over Time

The introduction of factor supply effects to empirical explanations of interest rates sheds new light on two other puzzles in the interest rate literature. Cargill and Carlson both argue that the interest rate–inflation rate relation has been unstable over time. When we omit *SUPPLY* from (6), the estimated expected inflation coefficient does vary substantially over different subsamples. Formal stability tests confirm this. The *F*-tests consistently reject the stability of that coefficient when *SUPPLY* is not included.<sup>27</sup> The last two rows of Table 2 present the results of estimating (6) when the sample is split at its midpoint. The point estimates for each coefficient vary little across subsamples.

<sup>27</sup>The data were all transformed by the estimated quasi-differencing coefficient from row 2 of 0.63. Instability results when the sample is split at the end of the 1950's, the end of the 1960's, or at the sample midpoint.

And contrary to the non-*SUPPLY* specification, we cannot reject the hypothesis of jointly stable coefficients, regardless of whether the sample is split at its midpoint, at the end of the 1950's, or at the end of the 1960's.<sup>28</sup>

Though we cannot reject the hypothesis that it was stable, it is worth noting that the point estimate for expected inflation is slightly higher (1.11 vs. 1.07) later in the sample. This runs contrary to Carlson's evidence that extending the sample to include the more recent period causes that effect to drop sharply. If we omit *SUPPLY*, however, and allow the expected inflation coefficient to shift in the 1970's, it does fall by over one-quarter, from 1.19 to 0.86. That apparent decline is explained by the change over time in the correlation between expected inflation and the omitted supply shock proxy. From 1952 to 1969, their correlation is  $-0.86$ ; from 1970 to 1979, it is  $+0.83$ . Thus omission of supply forces biases the estimated expected inflation effect upward before 1970 and downward after 1970, creating the illusion of a falling coefficient.

Another test of the stability of the interest rate relation is provided by letting the expected inflation coefficient follow a first- or second-degree polynomial in time. This specification allows for gradual evolution of the parameter  $\beta_1$  through time as opposed to the sharper break suggested by split-sample tests. In no case was there any evidence of instability; none of the time coefficients even approached conventional significance levels. Thus, our model leads to estimates that are not only significant but stable over a wide range of experience.

#### V. Conclusion

Real interest rates were dramatically lower in the 1970's than in the two previous decades. Rather than appealing to irrationality or coefficient instability, I hypothesize a link between real supply conditions and real returns. I argue that real interest rates fell in the latter 1970's in response to a reduction in

<sup>28</sup>All data were transformed by the quasi-differencing coefficient from row 3 of 0.36.

the supply of complementary factors of production, especially energy. As input prices rose, the profitability of and demand for capital fell. The lowered growth rate of the capital stock and concomitant decline of investment dragged down the real rate of interest. The estimates in Section II suggest that, by 1978, supply forces had pulled real pre-tax interest rates down 1.7 percentage points from their 1972 level. Corroborating evidence for a drop of this size is provided by Michael Bruno. His estimates of the factor price frontier for the U.S. manufacturing sector imply that increases in the relative price of material inputs drove down the after-corporate-tax profit rate by 1.5 percentage points.<sup>29</sup>

The estimated impact of such a shock proves to be economically significant and statistically robust with respect to sample period, to various specifications of the remaining independent variables, and to various specifications of the supply variable itself. In addition to lending support to the supply-shock hypothesis, the results go a long way toward eliminating the puzzling magnitude and instability of previous interest-inflation rate estimates. Including supply forces not only removes previously extant coefficient biases, but also yields an estimated interest rate function that is stable over the entire post-Accord period.

The substantial changes wrought by deregulation of the financial sector since 1979 are of the nature and magnitude that engender changes in structural (and reduced-form) coefficients. To the extent competition for funds has been sharpened, we would expect real rates to rise above predictions generated by models based on historical data, including this one. Nonetheless, my model would predict that expansionary fiscal policy, coupled with a reduction in the long-run money growth rate that depressed the expected inflation rate, would raise real after-tax inter-

est rates. Over the shorter haul, lower money growth would raise real rates even further. Increases in the supply of energy would likewise tend to raise real rates. To the extent others' forecasts omit this factor, however, my model predicts that real rates would rise above their predictions.

## REFERENCES

- Berndt, Ernst and Khaled, Mohammed, "Parametric Productivity Measurement and Choice Among Flexible Functional Forms," *Journal of Political Economy*, December 1979, 87, 1220-45.
- \_\_\_\_\_ and Wood, David O., "Engineering and Econometric Interpretations of Energy-Capital Complementarity," *American Economic Review*, June 1979, 69, 342-54.
- Bruno, Michael, "Raw Materials, Profits, and the Productivity Slowdown," Working Paper No. 660R, National Bureau of Economic Research, December 1981.
- Cargill, Thomas, "Anticipated Price Changes and Nominal Interest Rates in the 1950s," *Review of Economics and Statistics*, August 1976, 58, 364-67.
- \_\_\_\_\_ and Meyer, Robert A., "The Term Structure of Inflationary Expectations and Market Efficiency," *Journal of Finance*, March 1980, 35, 57-70.
- Carlson, John, "Expected Inflation and Interest Rates," *Economic Inquiry*, October 1979, 17, 597-608.
- Clark, Peter K., "Investment in the 1970s: Theory, Performance, and Prediction," *Brookings Papers on Economic Activity*, 1:1979, 73-113.
- Darby, Michael, "The Financial and Tax Effects of Monetary Policy on Interest Rates," *Economic Inquiry*, June 1975, 13, 266-76.
- Eckstein, Otto, "The Decline in Capital Formation in the Industrial World," in Allen R. Sanderson, ed., *DRI Readings in Macroeconomics*, New York: McGraw-Hill, 1981, 137-45.
- Fama, Eugene, "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, June 1975, 65, 269-82.

<sup>29</sup>This decline is calculated as the product of the after-profits-tax real return on manufacturing capital (7.84 percent, from Holland and Myers), the impact of relative input prices on profit rates (-1.526, from Bruno's fn. 18), and the change in the log of the ratio of materials prices to wholesale prices (0.122, 1982 *Economic Report of the President*, Tables B-55 and B-56).

- \_\_\_\_\_, "Interest Rates and Inflation: The Message in the Entrails," *American Economic Review*, June 1977, 67, 487-96.
- Feldstein, Martin, "Inflation, Income Taxes, and the Rate of Interest: A Theoretical Analysis," *American Economic Review*, December 1976, 66, 809-20.
- Friedman, Milton, "Factors Affecting the Level of Interest," in Donald Jacobs and Richard Pratt, eds., *Saving and Residential Financing: 1968 Conference Proceedings*, Chicago: U.S. Savings and Loan League, 1968.
- Gibson, William, "Interest Rates and Inflationary Expectations: New Evidence," *American Economic Review*, December 1972, 62, 854-65.
- Gordon, Robert J., "Alternative Responses of Policy to External Shocks," *Brookings Papers on Economic Activity*, 1:1975, 183-204.
- \_\_\_\_\_, "Comment," *Brookings Papers on Economic Activity*, 1:1980, 249-57.
- Gramlich, Edward, "Macro Policy Responses to Price Shocks," *Brookings Papers on Economic Activity*, 1:1979, 125-66.
- Holland, Daniel and Myers, Stewart, "Profitability and Capital Costs for Manufacturing Corporations and All Nonfinancial Corporations," *American Economic Review Proceedings*, May 1980, 70, 320-25.
- Hudson, Edward and Jorgenson, Dale, "Energy Prices and the U.S. Economy, 1972-1976," *Natural Resources Journal*, October 1978, 18, 877-97.
- Johansen, Leif, *Production Functions*, Amsterdam: North-Holland, 1972.
- Kopcke, Richard, "The Behavior of Investment Spending during the Recession and Recovery, 1973-1976," *New England Economic Review*, November/December 1977, 5-41.
- Levi, Maurice and Makin, John, "Anticipated Inflation and Interest Rates," *American Economic Review*, December 1978, 68, 801-12.
- Livingston, Joseph A., Surveys published twice yearly, *Philadelphia Sunday Bulletin*, 1948-71; *Philadelphia Inquirer*, 1972 on.
- Mishkin, Frederic, "The Real Rate of Interest: An Empirical Investigation," *Carnegie-Rochester Conference Series on Public Policy: Supply Shocks, Incentives, and National Wealth*, Spring 1981, 14, 151-200.
- Mundell, Robert, "Inflation and Real Interest," *Journal of Political Economy*, June 1963, 71, 280-83.
- Phelps, Edmund, "Commodity-Supply Shock and Full-Employment Monetary Policy," *Journal of Money, Credit, and Banking*, May 1978, 10, 206-21.
- Pyle, David, "Observed Price Expectations and Interest Rates," *Review of Economics and Statistics*, August 1972, 54, 275-80.
- Rasche, Robert, and Tatom, John, "Energy Resources and Potential GNP," *Federal Reserve Bank of St. Louis Review*, June 1977, 59, 10-23.
- Sachs, Jeffrey, "The Current Account and Macroeconomic Adjustment," *Brookings Papers on Economic Activity*, 1:1981, 201-68.
- Tanzi, Vito, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates," *American Economic Review*, March 1980, 70, 12-21.
- Tatom, John, "Energy Prices and Capital Formation: 1972-1977," *Federal Reserve Bank of St. Louis Review*, May 1979, 61, 2-11.
- Tobin, James, "Money and Economic Growth," *Econometrica*, 33, October 1965, 671-84.
- Wilcox, James A., "Interest Rates, Expected Inflation, and Supply Shocks, Or Why Real Interest Rates Were So Low in the 1970s," National Bureau of Economic Research Conference Paper No. 121, May 1981.
- International Monetary Fund, *International Financial Statistics*, 1980 Yearbook.
- U.S. Board of Governors of the Federal Reserve System, *Federal Reserve Bulletin*, Washington: USGPO, various years.
- \_\_\_\_\_, *Statistical Releases H.6 and G.13*, Washington: USGPO.
- U.S. Council of Economic Advisors, *Economic Report of the President*, Washington: USGPO, various issues.
- U.S. Department of Commerce, Bureau of Economic Analysis, *The National Income and Product Accounts of the United States*, Washington: USGPO.
- U.S. Department of Labor, Bureau of Labor Statistics, *Monthly Labor Review*, Washington: USGPO, various years.

# Loyalty Filters

By GEORGE A. AKERLOF\*

When people go through experiences, frequently their loyalties, or their values, change. I call these value-changing experiences "loyalty filters." This paper considers the case where these values are partially, but not totally, changeable. In addition, persons, by having a choice over their experiences, can exercise some choice over their values; or perhaps more typically, persons may choose for their children experiences that will lead them to have desired values. Insofar as this occurs, values are not fixed, as in standard economics, but are a matter of choice. Economic theory, which is largely a theory of choice, then becomes a useful tool in analyzing how these values are chosen. Most persons attempt to choose values for their children (and perhaps also for themselves) according to their economic opportunities that allow them to get along economically. According to Robert Coles' *Children of Crisis*, not only the wealthy (who will be discussed at some length in Section II), but also the poorest of the poor—immigrants, sharecroppers, and mountaineers—consciously teach their children values aimed at leading them best to survive economically.

*The Wealth of Nations* concerned itself with the issue of how the economy would behave if everyone were to behave selfishly. Adam Smith's famous answer to this question in terms of the invisible hand is the key result in economic theory. Since the time of Edgeworth (see Amartya Sen, 1977, p. 317), it has been fashionable for non-Marxist economic theorists to follow Smith's presumed worst-case assumption—that all persons are totally selfish. Yet as Sen points out, this assump-

tion is made for reasons of convenience, not because economists empirically assume that all persons act only out of selfishness.

This paper will explore the extent to which parents interested only in their children's economic welfare will teach them to be totally selfish. Section I gives an example in which children are taught to be honest, even to their own detriment. Such a teaching may cause children to act against their own short-run interest even while it serves their long-run economic interests. Similarly, Section II yields a model where children are taught to be loyal to their class interests; this teaching may not serve their individual short-run interest, but it does serve their individual long-run interest. Each of these models is motivated by an empirical observation. In the case of Section I, this observation concerns the economic well-being of the high-minded Quakers: if selfishness pays off, why should the Quakers do so well? Section II is motivated by Coles' studies of the way in which privileged children learn to view those less fortunate as "others," in contrast to "us."

The models of Sections I and II are meant to show economic man as not being undeviatingly selfish. In Section I, he is undeviatingly honest, even against his interest, and in Section II he is undeviatingly loyal to his class interest. Yet at the same time his long-run interests have been maximized by teaching him a code of conduct that leads him, insofar as possible, to act in his best long-run interest. Section III continues the process of making economic man less undeviatingly selfish. This section concerns economic and political elites who are the products of consciously styled elite academies. Examples of such elites come from military service academies, prestige universities, and other institutions that not only give technical training, but also teach loyalty to these institutions and the type of persons who are their faculty or alumni. Where these institutions are aligned with the government (or else

\*University of California-Berkeley. I thank Donald Hayes, Hajime Miyazaki, and Janet Yellen for invaluable help and comments, and the Institute of Business and Economic Research, University of California-Berkeley, for logistical support. I also thank the National Science Foundation for financial support under research grant no. SES-8119150.

where their graduates have other monopoly powers), it is shown that the curriculum that best serves its alumni not only teaches technical skills, but also loyalty to the type of person who is a fellow graduate. In the model of this section, the elite graduate is unselfish in serving his country; nevertheless, due to the biases in his values, the interests of the elite end up being served, as well perhaps as the interest of the country. The picture emerges of well-trained, well-meaning civil servants who act selflessly according to their best conscience, yet nevertheless manage to earn more than the competitive wage due to the cultural biases that have been chosen.

Finally, before getting into the specific models, I would like to make a few remarks. Albert Hirschman's *Exit, Voice and Loyalty* (1970) is the only recent non-Marxist economics to emphasize the role of loyalty in economic theory.<sup>1</sup> Yet, for the most part, his book is unconcerned with how loyalties begin, which is the focus of this paper. I resisted the temptation to call this paper "Entrance, Voice, and Loyalty," which would have emphasized the contrast with Hirschman's work, because the title "Loyalty Filters" better conveys the generality of my subject matter. In this regard, I would like to remark on the particular and illustrative nature of the examples that follow. They fail in their particularity to reflect the many important possible types of loyalty filter. The agent who experiences the filter may consciously or unconsciously choose the experience. He may be conscious or unconscious of the effect of the experience on his loyalties. And the experience may not only be chosen by himself (or an agent such as his parents acting on his behalf), but instead by another

agent acting in his own selfish interest, such as an advertiser interested in fostering brand loyalty to the product he sells, or an employer interested in extracting unselfish performances from his employees.<sup>2</sup> Furthermore, according to George Homans (1950), loyalties change according to almost every role a person plays and almost every situation that involves him. The preceding rudimentary classification of loyalty filters according to choice/consciousness/agent choosing/role of agent should alert us to the great variety of loyalty filters. The examples given below are meant as an illustrative teaching device and as an invitation to the reader to roll his own examples of loyalty filters.

## I. A Model of Honesty and Cooperative Behavior

### A. Motivation

As mentioned above, the model in this section is motivated by the assumption of selfishness in economic models. It is also motivated by an experiment in social psychology (Fred Arnstein and Kenneth Feigenbaum, 1967). In this experiment, persons of different religious persuasions were asked to play a game of the prisoner's dilemma variety; in this game, noncooperative behavior improved considerably the lot of the noncooperative player provided the other player's behavior remained cooperative. Conversely, cooperative players fared quite poorly if the other players were noncooperative. It turned out that the Quakers, as might be expected, ranked quite high in terms of the trustfulness and cooperation of their responses, but low in terms of their economic rationality. This result is curious because in real life, Quakers are usually considered one of the wealthiest minority groups in the United States (Gordon Allport, 1958, p. 72). The model below is intended, accordingly, to show that honesty and cooperative behavior pay off; the honest person is not just a systematic "sucker."

<sup>1</sup>In Marxist terminology, this paper concerns how a class "in itself" becomes a class "for itself" (Anthony Giddens, 1975, p. 30). The prediction following Marx that most poverty stricken in society will be reactionary in attitude (Giddens, p. 37) accords exactly with that of the model of class loyalty in Section II. Only extra-economic attitudes will cause the poor to unite in their own class interests according to that model. That prediction is also consistent with the observation that most socialist revolutions have occurred in the wake of wars fought for reasons only incidental to the socialist takeover which later occurs.

<sup>2</sup>A very different type of loyalty filter from those in this paper is analyzed by myself and William Dickens (1982). I would like to record my debt to him for what I learned while jointly writing that paper.

The model is quite trivial; it corresponds exactly to a known observation: couriers who carry large sums of money are often "bonded." Apparently, it pays to bond such persons, which in effect is to guarantee their honesty. According to the model here, it pays persons to bond themselves by acquiring traits that cause them to appear honest. And the cheapest way to acquire such traits according to our model is, in fact, to be honest! This distinction between *appearance* and *actuality* of honesty has been discussed by Max Weber (1958).<sup>3</sup> Weber's essays on the Protestant ethic are the classic description of how different experiences result in different personality types, with important economic consequences.

### B. The Model

*The Nature of Jobs* (from which labor demand is derived): Let there be only one type of job in the economy and let this job have a product  $y$ . Workers in this job, however, have an opportunity to embezzle an amount  $x$  with probability  $q$  unless there is surveillance. It would be straightforward to let there be surveillance costs, but complication is avoided by assuming these costs to be prohibitive.

*The Nature of Workers* (from which labor supply is derived): The utility of a worker depends on his income according to the utility

function  $u(\cdot)$ . Parents wish to maximize their children's welfare. They can train their children to be dishonest, which in this model means to embezzle whenever they can get away with it; or they can train their children to be honest.

*Equilibrium Wages of Dishonest Workers:* Employers are not fooled about the characters of their employees. A dishonest worker will be seen as such and his wage will be reduced by his expected embezzlement. Assuming risk-neutral competitive employers, a dishonest worker will receive a wage of  $y - qx$ , which is his product net of his expected embezzlement. Remember, however, that the worker has a chance to embezzle  $x$  with probability  $q$ . Thus with chance  $(1 - q)$ , the dishonest worker has utility  $u(y - qx)$ , and with chance  $q$ , he has utility  $u(y - qx + x)$ . The net result is an expected utility given by

$$(1) \quad E(u) = (1 - q)u(y - qx) + qu(y + (1 - q)x).$$

*Equilibrium Wage of Honest Workers:* Alternatively, according to the model, parents may train their children to be honest. Such training may require a cost to the parents, which we assume to be paid by the children. We call this cost  $c_h$ . Employers will pay honest persons their product  $y$ , so that their income net of training costs is  $y - c_h$ , and their utility is  $u(y - c_h)$ . Parents interested in maximizing their children's welfare will choose to make their children honest provided

$$(2) \quad u(y - c_h) > (1 - q)u(y - qx) + qu(y + (1 - q)x).$$

This last inequality always holds for given  $q$ ,  $y$ , and  $x$ , provided  $c_h$  is sufficiently small and  $u$  has diminishing returns.

*Very Dishonest Behavior:* There is a final question. Children could presumably also be taught to act honest yet embezzle when they get a chance. I will assume that such training is quite costly. Suppose the cost of such training is  $c_{od}$  (vd for very dishonest) and  $c_{od} - c_h > qx$ . In this case, the costs of such

<sup>3</sup>The coincidence between Weber's view and that of this paper regarding honesty can be seen in the following discussion of Benjamin Franklin by Weber:

Now all Franklin's moral attitudes are coloured with utilitarianism. Honesty is useful, because it assures credit; so are punctuality, industry, frugality, and that is the reason they are virtues. A logical deduction from this would be that where, for instance, the appearance of honesty serves the same purpose that would suffice, and an unnecessary surplus of this virtue would evidently appear to Franklin's eyes as unproductive waste.... But in fact the matter is not by any means so simple. Benjamin Franklin's own character, as it appears in the unusual candidness of his autobiography belies that suspicion. The circumstance that he ascribes his recognition of the utility of virtue to a divine revelation which was intended to lead him in the path of righteousness shows that something more than mere garnishing for purely egocentric motives is involved. [pp. 52-53]

In the view of this paper, Franklin, no matter how utilitarian his beliefs, could not acquire the appearance of honesty without its actuality.

training exceed the gains from dishonesty  $qx$ , so that it never pays parents to train their children in this way. It may appear at first glance that it should not be difficult to teach people to dissemble their values. But persons often have a hard time hiding their true nature. The rareness of acting talent can be perceived any day of the week by a comparison of daytime and nighttime TV.

Furthermore, there is evidence that traits once acquired (in this case, honesty) are often difficult to lose even when they have become dysfunctional. Robert Merton describes how bureaucrats' "adherence to the rules, originally conceived as a means, becomes transformed into an end-in-itself" (1956, p. 253). In my model of childrearing, honesty may begin as a means for economic betterment, but then there is a displacement of goals so that the person so trained will refrain from embezzlement where there is no penalty. Psychological experiments with animals show similarly that animals may quite easily be trained to have dysfunctional behavior. See, for example, Henry Gleitman's example (1981, p. 148) of trained helplessness in dogs.

*Remarks:* The role of jointness of production in training in this model should be noted. It is assumed that at the cost  $c_h$ , parents can train their children to *appear* honest. But to make children appear honest, it is easiest to make them also *be* honest. There is a return to *appearing* honest, but not to *being* honest. It pays parents to teach their children to be honest because the individually functional trait of appearing honest is jointly produced with the individually dysfunctional trait of being honest.

It should also be noted that the word *embezzlement* need not be taken too literally. Any form of noncooperative behavior by workers, which the firm will find expensive and difficult to police, can play the exact same role as embezzlement. In many jobs, workers are given considerable scope for lack of cooperation before there will be retaliation by their supervisor. (See my 1982 article for an earlier discussion.) Such lack of cooperation has consequences similar to embezzlement in the model which has been presented. In the next section, a very similar model is proposed; embezzlement, however,

is replaced by the more general concept of noncooperation.

## II. Class Loyalty

This section concerns a theory of class loyalty and formation. Although individualistic economic theory is based upon the assumption that individuals act selfishly out of their own interests, it is certainly empirically and theoretically possible that persons are loyal to other ideals. Coles' *Privileged Ones* describes how wealthy children think about poorer persons. The model in this section is motivated by his study.

### A. *Studies of Social Identification* (Coles, Allport)

Coles' books, which are the result of fifteen years' intensive work by a trained child psychologist, may be the most complete and detailed study ever made of the formation of class loyalties. Nevertheless, the whole process whereby the socialization occurs is still a bit mysterious, even from Coles' detailed accounts. Why this mystery is of necessity the case is explained by Allport in his discussion of the learning of social values via identification of children with their parents:

Learning through identification seems basically to involve a type of muscle strain or postural imitation. Supposing the child, hypersensitive to parental cues, senses a tightness or rigidity when his parents are talking about the Italian family that has moved in next door. In the very act of perceiving these parental cues, the child grows tight and rigid.... After this associated experience, he may tend, ever so slightly, to feel a tenseness (an incipient anxiety) whenever he hears (or thinks) of Italians. The process is infinitely subtle. [1958, pp. 278-79]

Despite this subtlety, because of the intensity of his study, Coles is able to report, here and there, the emergence of social values among the wealthy as the younger children in his sample ask their mommy or daddy

embarrassing questions such as why their family should not share their wealth with others. There are many answers to these questions, such as "Daddy works hard for what he earns"; "Mommy and Daddy give a great deal to charity"; "there are so many poor persons our contribution could only be very small." In some cases, the questions persist, particularly where the children identify with the maid and possibly also with her children. However, in these cases of persistence, these questions are usually abandoned when mommy (typically) makes it clear that it is not nice to annoy the loving daddy with such annoying and persistent questions. The children then re-immerses themselves in leading the "busy, busy lives" (Coles' phrase) which their parents have planned for them.

Although much of the process of socialization is difficult to see, it is clear that it is quite intentional on the part of these wealthy parents that their children are taught to view themselves as "different" from those who are less fortunate. This does not mean, typically, that "others" are to be despised; but almost all the wealthy children in Coles' book have a sense of identification with "us," children and families who are equally well to do, in contrast to "others" who are less fortunate.

The role of the difference in the lives of rich children and poor children in causing this sense of distinction and identification is made clear in the description of a young New Orleans girl at the time of the racial trouble in the early 1960's. I wish to stress, as does Coles, that the social meanings ascribed by this girl, although seen through the eyes of a child, are, nevertheless, exactly those intended by her parents. According to Coles, "'Our maid's children don't know about finger bowls,' a seven-year-old New Orleans girl says. She also says—the year is 1960—that 'the kids going into those [desegregated] schools don't know about finger bowls either; and they don't know how to smile and say thank-you to the people in the mobs'" (1977, p. 530).

The girl's plan of action for the black children going into the desegregated schools is to be unfailingly polite (i.e., to learn about finger bowls). This plan of action and the

statement I have quoted are explained by Coles as correctly perceiving the role of manners in differentiating the wealthy from the poor. Displaying such politeness, in the view of the child, the black children will achieve her own status (and that of her family) and the mobs will cease to be hostile.

To summarize, wealthy parents tend to teach their children an identification with other persons of wealth and to view the less fortunate as others. Furthermore, this teaching is quite intentional, either learned through certain coded messages such as "manners" inherent in the way of life of the wealthy, in subtle demonstrations of annoyance or tension by parents, or, finally, occasionally, but except in rare cases only to younger children, by requests not to annoy mommy or daddy with needless questions. This section will construct a model wherein parents interested in maximizing their children's economic welfare will teach them such values. According to the technology, parents may teach their children to have such class values, but such values, as in the earlier model of honesty, cannot be dissembled. Thus persons cannot pretend to identify with other members of their class without actually being loyal.

### B. A Model

The model will be quite similar to the one in the last section, only with some added generality. Let there be two types of persons, the wealthy, represented by  $W$ , and the poor, represented by  $P$ . Suppose that a  $W$  may be hired either by a  $P$  or a  $W$ , and that a  $W$  will have a marginal product, if cooperative, in a job provided by a  $W$  of  $y_{WW}$  and will have a chance of engaging in noncooperative behavior that will reduce this marginal product by  $x_{WW}$  with probability  $q_{WW}$ . The double subscript  $WW$  indicates that a  $W$  (the first subscript) is providing a job to a  $W$  (the second subscript). The probability that a  $W$  so hired will engage in noncooperative behavior will depend on his class loyalties. If loyal to the  $W$ , his probability of noncooperative behavior will be low; if loyal to the  $P$ , his probability of noncooperative behavior in such a job will be high. Again, as in the previous model,

and again only for modeling convenience, assume that surveillance is not possible.

In the usual neoclassical model, contracts may be made between any two persons. Assume, therefore, similarly and symmetrically with the  $WW$  case, that a  $P$  may also hire a  $W$ . The marginal product of a  $W$  working for a  $P$  is  $y_{PW}$  with a chance of noncooperative behavior  $q_{PW}$  with cost to the employer of  $x_{PW}$ .

Consistent with the notation of the earlier model, the cost of instilling class loyalties is denoted by the letter  $c$ , with  $c_{WW}$  the cost of instilling loyalty of a  $W$  to the  $W$  and  $c_{PW}$  the cost of instilling loyalties of a  $W$  to the  $P$ .

The individual person maximizes his utility. This utility depends on his income and also on his behavior on the job, which may be cooperative or noncooperative. The individual benefits from noncooperative behavior, but it is not automatic that such behavior that costs the employer  $x_{WW}$  or  $x_{PW}$  will result in a benefit to the employee of equal amount. I will assume that the individual values the returns from noncooperative behavior at a fraction  $\alpha$ ,  $0 \leq \alpha \leq 1$ , of its cost. As before, let the individual have a utility function  $u(\cdot)$ , which in this case depends on the wage plus the value to him of being noncooperative, if he chooses that mode of behavior.

The individual has three choices: whether to work for a  $W$  or a  $P$ ; whether to be loyal to the  $W$  or not; whether to be loyal to the  $P$  or not. In general, a person with a chance of noncooperation  $q$  at cost  $x$  in a job with cooperative product  $y$ , and with loyalties that are acquired at cost  $c$ , will receive a wage  $y - qx$  and therefore have an expected utility

$$(3) \quad E(u) = (1 - q)u(y - qx - c) + qu(y - qx + \alpha x - c).$$

Adoption of the following notation allows a single expression for the maximization problem of a  $W$ . Let  $e_{WW}$  be a dummy variable equal to unity if  $W$  is employed by a  $W$ , and equal to zero otherwise; let  $l_{WW}$  be a dummy variable equal to unity if  $W$  is loyal to the  $W$ , and equal to zero otherwise; and let  $l_{PW}$

be a dummy variable, similarly, equal to unity if  $W$  is loyal to the  $P$  and equal to zero otherwise.

Accordingly, a  $W$  chooses for himself (or for his child) the variables  $e_{WW}$ ,  $l_{WW}$ ,  $l_{PW}$  to maximize  $E(u)$ , which is given by the expression:

$$(4) \quad E(u) = e_{WW}E_{WW}(u) + (1 - e_{WW})E_{PW}(u),$$

where  $E_{WW}(u)$  and  $E_{PW}(u)$  are the expected utilities of a  $W$  working for a  $W$  and a  $P$ , respectively. (An explicit expression for  $E(u)$  can be derived as a function of  $e_{WW}$ ,  $l_{WW}$ , and  $l_{PW}$  by use of (3). Let  $E_{WW}$  and  $E_{PW}$  be written as functions of  $l_{WW}$  and  $l_{PW}$  by insertion into (3) of the appropriate subscripts on  $E(u)$ ,  $q$ ,  $y$ ,  $x$ , and  $c$ , with  $q_{WW}$  and  $q_{PW}$  each an explicit function of its two arguments,  $l_{WW}$  and  $l_{PW}$ . Substitution of the expressions for  $E_{WW}$  and  $E_{PW}$  into (4) yields  $E(u)$  as a function of the three optimizing variables,  $e_{WW}$ ,  $l_{WW}$ , and  $l_{PW}$ .)

Assume that there are  $K_W$  units of capital owned by the  $W$  and  $K_P$  units of capital owned by the poor, and that both types of capital use both  $W$  and  $P$  labor with constant returns to scale. Then in equilibrium  $E_{WW}(u) = E_{PW}(u)$ , or, in words, the expected utility of a  $W$  working for a  $W$  and for a  $P$  are equal. The number of  $W$  working for  $W$ -capital and the number of  $W$  working for  $P$ -capital will be proportional to  $K_W$  and  $K_P$  respectively; thus the fraction of  $W$  working for  $W$  is  $K_W / (K_P + K_W)$ .<sup>4</sup>

<sup>4</sup>This intuitive result depends upon various assumptions in addition to constant returns to scale. The production functions with  $W$  and  $P$  owners of capital must be the same. In addition the following symmetry conditions are required:  $q_{WW}(x, y) = q_{PW}(y, x)$ ,  $q_{WP}(x, y) = q_{PP}(y, x)$ ,  $x_{WW} = x_{PW}$ ,  $x_{PP} = x_{WP}$ ,  $c_{WW} = c_{PW}$ , and  $c_{PP} = c_{WP}$ . These conditions guarantee that if the capital/ $W$ -labor and capital/ $P$ -labor ratios are the same with  $K_W$  and  $K_P$ , respectively, the expected marginal products and utilities of both  $W$ -labor (and also  $P$ -labor) will be equal on  $W$ -capital and on  $P$ -capital. The equilibrium condition that  $E_{WW}(u) = E_{PW}(u)$  requires, of course, that some  $W$  are working for  $P$ , as well as  $W$ , capital. Despite the necessity of the stringent assumptions to show that the number of  $W$  working for  $W$ -capital is exactly proportional to  $K_W$ , the result that most workers will be working for  $W$ -capital if most capital is owned by  $W$ , will be quite robust.

In equilibrium, a  $W$  is indifferent to working for a  $W$  or a  $P$ , but if working for a  $W$ , the maximizer will choose loyalties to the  $W$  and not to the  $P$ ; if working for the  $P$ , the opposite choices will be made, provided  $c_{WW}$  and  $c_{PW}$  are sufficiently small.

A  $W$  working for a  $W$  will receive expected utility which can be expressed (with appropriate use of subscripts in (3)) as

$$(5) \quad (1 - q_{WW}(l_{WW}, l_{PW})) \\ u(y_{WW} - q_{WW}(l_{WW}, l_{PW})x_{WW} \\ - l_{WW}c_{WW} - l_{PW}c_{PW}) + q_{WW}(l_{WW}, l_{PW}) \\ u(y_{WW} - (q_{WW}(l_{WW}, l_{PW}) - \alpha)x_{WW} \\ - l_{WW}c_{WW} - l_{PW}c_{PW}).$$

Equation (5) is maximized (provided  $c_{WW}$  is sufficiently small, and given the assumptions that  $q_{WW}$  decreases with  $l_{WW}$  and increases with  $l_{PW}$ ), by choosing loyalty to the  $W$  and no loyalty to the poor. In mathematical terms, this means choosing  $l_{WW}$  equal to unity,  $l_{PW}$  equal to zero.

Since most nonresidential capital is either owned or controlled by the wealthy, it may be assumed that  $K_W$  is large relative to  $K_P$  and hence it pays most  $W$  to train their children to be loyal to the  $W$  and not to the  $P$ .

### C. Loyalties of the Poor

The previous model agrees with empirical findings regarding the loyalties of the wealthy as described by Coles. How does it fare with respect to its predictions regarding the loyalties of the poor?

Insofar as the assumptions of the stringent model hold, it predicts that the poor will also be loyal to the wealthy in proportion  $K_W/(K_P + K_W)$ . This is consistent with Coles' findings regarding the teaching of poor mothers to their children, although this loyalty may be instilled more out of passive acceptance of the system than out of genuine enthusiasm—as suggested by the words of one very poor migrant mother: “Do you

have a choice but to accept?... Once, when I was little I seem to recall asking my uncle if there wasn't something you could do, but he said no, there wasn't and hush up. Now, I have to tell my kids the same, that you don't go around complaining—you just don't” (Coles, 1967, p. 52).

Of course there are cases that run counter to the predictions of the model, where the poor have not been loyal to those who provide them with jobs. The model, in fact, gives predictions where such conditions are likely to occur because its assumptions are violated. Where willing cooperation is not necessary from workers, it is not necessary to secure their active loyalty. Such devices as the assembly line force workers to work at the pace of the line irrespective of their mental attitude. Also, such incentive schemes as piecework, where the worker who puts in less effort receives correspondingly less pay, reduce the cost to the employer of unwilling workers, and therefore causes there to be less reason why firms should demand positive loyalty. In contrast, servants' cooperative willingness is often of positive value; and, correspondingly, the term “loyal servant” is a standard figure of speech.

Furthermore, social institutions may change the loyalty incentive structure. As one example, unions that interpose themselves between workers and the firm regarding work conditions reduce the positive incentives for workers to be cooperative. Welfare is another agency that reduces the incentives for parents to teach their children traditions of cooperation with employers, either rich or poor.

While this model is all too simple to predict class loyalties in many complicated situations, particularly where feelings of justice and fairness play an important role, nevertheless, in capturing some of the economic incentives for being cooperative (vs. noncooperative), this model does allow some comparative static analysis of class loyalties. At the minimum it serves as a reminder that an important side effect of social policy (toward unions and welfare, for example) is the resultant change in loyalties due to the change in incentive structure.

### III. Institutional Loyalties

The preceding sections have shown that parents eager to maximize their children's economic welfare may find it advantageous to teach honesty (Section I) and class loyalties (Section II) even though these traits may in some circumstances cause the individual to engage in nonmaximizing behavior. It pays parents to teach honesty and class loyalty because the *appearance* of honesty and class loyalty are beneficial; the easiest way to achieve these appearances is to *be* honest and loyal, even though honesty and loyalty themselves involve sacrifices.

This section presents a similar model of elite institutions. According to this model patriotism is jointly taught (i.e., jointly produced) with cultural values that are favorable to fellow graduates of the institution. This jointness of loyalties is reflected in the statement of President Eisenhower's defense secretary: "For years I have thought what was good for our country was good for General Motors and vice versa" (*New York Times*, 1954); likewise it is reflected in the college song which ends "for God, for country, and for Yale."

According to my model, loyalty to the institution has the effect that the services of graduates of the institution are highly valued—indeed, overvalued—by other graduates. Thus, while graduates of the institution may be patriotic even to the point of considerable self-sacrifice, the teaching of this patriotism may be of economic benefit to the graduates because it occurs jointly with the teaching of cultural biases in favor of the institution's graduates.

#### A. Examples

It may be helpful to give some concrete examples of institutions that, at least arguably, correspond to the model. In the United States, the military service academies teach loyalty to the academies themselves and also to the country. In Britain, Oxford and Cambridge—and, for some graduates, the public schools prior to university—teach loyalty to British values in general, and good govern-

ment in particular. In addition, there appears to be considerable loyalty to fellow graduates. As one indication of this loyalty, most MPs of both the Labour and Conservative parties are graduates of these two universities; more remarkable still, eighteen out of twenty-two of Mrs. Thatcher's current cabinet members are graduates of public schools, only two are nongraduates of Oxbridge or the Army-Navy service academies.

#### B. The Model

In this model, it is assumed that the public has a choice (as in fact may not actually occur) between government by elite-school graduates and nongraduates. It is assumed that the nongraduates are loyal to themselves (because they have not been taught the elite patriotism), while the elitists are patriotic but with elite biases. The public, interested in good government, chooses the elite-school graduates as ministers. These graduates are patriotic and self-sacrificing, but in such a way that the interests of graduates are served on the average.

Assume that there are two types of persons—graduates and nongraduates. The government needs ministers in number equal to a fraction  $\alpha$  of all graduates. These ministers, like other graduates, have a marginal product  $w$  outside the government. Graduates, being patriotic, are willing to serve as ministers for remuneration which is a fraction  $\beta$  of  $w$ . Ministers award government contracts. Ministers who are graduates value the services of other graduates at  $(1 + \gamma)w$ , whereas these services elsewhere only have value  $w$ . The fraction  $(1 - \alpha)$  of graduates who do not work as ministers are hired by the government as long as their wage paid there exceeds their marginal products elsewhere.  $\beta$  and  $\gamma$  are both functions of the curriculum, denoted  $c$ , of the elite institution.

Nongraduates have no cultural biases, but, by assumption, they do not have the patriotism of the elite. Consequently, they wish to award government contracts for their own benefit; by assumption, it is impossible to check on such misappropriation, and therefore the return to the public from non-

graduates is zero. The expected return from government contracts awarded by graduates to a graduate is  $w$  at a cost  $(1 + \gamma)w$ . Since the benefit-cost ratio of a government of graduates, even if not optimal, is higher than of nongraduates, the public chooses the former type of government.

Now consider the expected return to graduates which, if the economic-maximizing curriculum  $c$  is chosen, will be

$$(6) \max_c \{ \alpha \beta(c)w + (1 - \alpha)(1 + \gamma(c))w \}.$$

In the case of an internal equilibrium where curriculum is a continuous variable, the curriculum which is the economic optimum for graduates will meet the condition that the marginal decrease in wages due to self-sacrifice just balances the marginal return to other graduates, due to the overvaluing of their services. Note that the economic optimum for graduates in this model is not the economic optimum for the public. For the public, the best curriculum is one that *maximizes* the benefits net of costs of contracts, including the costs of hiring fellow graduates.

#### C. Summary

The following phenomenon has been modeled. Graduates of elitist institutions are often excellent at their jobs and genuinely interested in the "common welfare" as they see it. While they give less than the best possible service because of that elitism per se, however, that is better yet than what would be given by persons who remain untrained in values of patriotism and loyalty to the organization. The graduates on the average, although they sometimes do genuinely sacrificial service, still have a positive economic return, because what is lost due to their sacrifice is more than offset by the overvalue by the government in the award of government contracts. The net result yields less than the optimum to the nongraduate public; control of the elite curriculum, or other government regulations such as "affirmative action" hiring of nongraduates, will improve government benefit-cost performance.

#### IV. Conclusion

This paper has presented examples of the concept of loyalty filters and their potential importance for economic theory. According to the key idea underlying this paper, as persons go through different experiences, their loyalties change. Loyalty filters have implications for how individuals and institutions will attempt to reach specified goals, as illustrated above, where, in each of the three examples, the goal was the maximization of economic welfare. Loyalty filters, as well, have implications concerning the goals that individuals attempt to attain. The modeling of each of these aspects of reality constitutes a departure of importance from standard economic models, capable of explaining such phenomena as cooperative behavior, class loyalties, and much institutional behavior.

#### REFERENCES

- Akerlof, George A., "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, 2.
- \_\_\_\_\_ and Dickens, William T., "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, June 1982, 72, 307-19.
- Allport, Gordon W., *The Nature of Prejudice*, Anchor Books ed., Garden City: Doubleday and Company, Inc., 1958.
- Arnstein, Fred and Feigenbaum, Kenneth D., "Relationship of Three Motives to Choice in Prisoner's Dilemma," *Psychological Reports*, June 1967, 20, 751-55.
- Coles, Robert, *Migrants, Sharecroppers, Mountaineers*, Vol. II—*Children of Crisis*, Boston: Little, Brown and Company, 1967.
- \_\_\_\_\_, *Privileged Ones: The Well-Off and the Rich in America*, Vol. V—*Children of Crisis*, Boston: Little, Brown and Company, 1977.
- Giddens, Anthony, *The Class Structure of the Advanced Societies*, New York: Harper and Row, 1975.
- Gleitman, Henry, *Psychology*, New York: W. W. Norton, 1981.

Hirschman, Albert O., *Exit, Voice, and Loyalty*, Cambridge: Harvard University Press, 1970.

Homans, George C., *The Human Group*, New York: Harcourt, Brace & World, 1950.

Merton, Robert K., *Social Theory and Social Structure*, New York: The Free Press, 1956.

Sen, Amartya K., "Rational Fools: A Critique

of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs*, Summer 1977, 6, 317-44.

Weber, Max, *The Protestant Ethic and The Spirit of Capitalism*, New York: Scribner's, 1958.

*New York Times*, Section VI, 4:4, February 28, 1954.

# Property Rights, Transaction Costs, and X-Efficiency: An Essay in Economic Theory

By LOUIS DE ALESSI\*

Real and imagined limitations of neoclassical economic theory have stimulated a variety of proposed revisions and *ad hoc* models. Two of these revisions are particularly interesting because they share the same point of departure but proceed along different routes to drastically different prescriptions. One approach extends the utility-maximization hypothesis to all individual choices under constraints, taking account of institutional restrictions and transactions costs in addition to the usual constraints. This generalization began to gain momentum with work by Armen Alchian on property rights (1959, 1961, and, with Reuben Kessel, 1962), Oliver Williamson on managerial discretion (1963), and Ronald Coase on transaction costs (1960). Since then, various aspects of the analysis have been developed more rigorously (Eirik Furubotn and Svetozar Pejovich, 1972; Alchian and Harold Demsetz, 1972; Williamson, 1975, 1979; Benjamin Klein, 1980) and have been exposed to empirical tests with generally favorable results (my 1980 article). The second approach also focuses on the individual as the basic unit of analysis, but rejects maximizing behavior in favor of other auxiliary hypotheses. This is the X-efficiency framework which Harvey Leibenstein first proposed in 1966, and has since refined and presented in a variety of forums.<sup>1</sup>

Surprisingly, proponents of these two approaches have not yet drawn on each other's contributions. Thus, Leibenstein has consistently offered his X-efficiency construct as an alternative to *traditional* neoclassical the-

ory without taking into account the property rights/transaction costs generalization. Indeed, Leibenstein even failed to note the existence of the latter during his recent search for "micro-micro," a link he missed in economic theory (1979a).<sup>2</sup> At the same time, although Alchian (1965a) and others have shown that differences in property rights can explain evidence used to support the existence of X-efficiency, they have not explicitly addressed Leibenstein's theoretical framework. Because both approaches insist on the paramount importance of the individual as the basic unit of analysis and both seek to explain the same phenomena, it seems useful to assess their relative merits and to inquire whether Leibenstein's criticisms of neoclassical economics also apply to the generalized neoclassical theory.

Section I contains a brief summary of neoclassical economic theory and its major criticisms. Section II generalizes neoclassical theory to all choices under constraints, and Section III explores the meaning of efficiency. Sections IV, V, and VI examine the construct, implications, and evidence of X-efficiency.

## I. Neoclassical Theory and its Critics

According to neoclassical theory, the individual consumer maximizes a single valued, convex, twice-differentiable utility function subject to a budget constraint. The budget constraint is determined by the prices of the rights to the use of the (homogeneous) commodities in the individual's choice set and by income. Income, in turn, is determined by the quantity and by the (derived) prices of

\*University of Miami. I am indebted for research support to the Law and Economics Center, and for helpful criticisms to Armen Alchian, David G. Davies, Robert Staaf, Judy Miley, and an anonymous referee of this *Review*.

<sup>1</sup>See the Leibenstein citations given in the Reference section.

<sup>2</sup>Although Leibenstein acknowledges a paper by Alchian and Demsetz (1972) on the theory of the firm, he does not consider the property rights/transaction costs alternative on which their analysis rests.

the rights to the use of the (homogeneous) resources which the individual owns, including the fractional ownership of business firms. The individual typically is a price taker both as a buyer of (the rights to the use of) commodities and as a seller of (the rights to the use of) resources. The state of nature constrains the stock of resources, whose initial distribution is given, and the state of the arts constrains how business firms may convert resources into commodities. Production functions are convex, twice-differentiable, and eventually exhibit decreasing returns to scale. Each business firm maximizes profits subject to its demand and cost conditions. To derive and test implications of this theory, therefore, it is both necessary and sufficient to identify the variables which enter utility and profit functions, and to indicate how changes in constraints affect the appropriate opportunity sets.

Additional characteristics of neoclassical theory deserve emphasis. In particular, 1) transaction costs are zero: broadly interpreted, this means that the costs of obtaining information about alternatives and of negotiating, policing, and enforcing contracts are zero; 2) adjustment costs are zero; 3) all resources are fully allocated and privately held; 4) owners allocate resources to productive purposes purely in response to pecuniary incentives; and 5) the entrepreneur's choice between income and leisure is independent of income (Tibor Scitovsky, 1943).<sup>3</sup> Thus, shirking by owners and by employees (including managers) is ruled out, and the profits of business firms (i.e., owners' wealth) are maximized.

Equilibrium in a neoclassical world satisfies all the Pareto-efficiency conditions under both competitive and monopolistic market structures.<sup>4</sup> Given zero transaction costs and the other attributes of neoclassical theory,

each monopolist would be on its least-cost expansion path and would discriminate perfectly among consumers, selling each additional unit to each consumer at exactly the maximum price which that consumer would be willing to pay until, at the margin, price would equal marginal cost. Other institutional arrangements, such as consumers' co-operatives, would also yield marginal cost pricing.

The purpose of economic theory presumably is to predict (explain) how changes in circumstances affect economic behavior. The relationships asserted by neoclassical theory, however, like those of most scientific theories, concern the behavior of idealized variables under highly purified conditions (for example, see Richard Braithwaite, 1960, p. 2). Accordingly, additional hypotheses are needed to relate the theory to real world phenomena (Ernest Nagel, 1963).

Dissatisfaction with the neoclassical framework focused largely on the theory of the firm, especially on the hypotheses that firms maximize profits and produce on the least-cost expansion path. Beginning in the late 1950's, economists proposed a variety of reforms most of which emphasized *ad hoc* models. Subject to some profit constraint, managers were hypothesized to maximize such things as the rate of growth of the firm (Edith Penrose, 1959), sales (William Baumol, 1959), the rate of growth of sales (Baumol, 1962), and the size of the firm (Robin Marris, 1964). In these and similar formulations (Milton Kafoglis, 1969), however, no criteria were provided for determining *a priori* the level of the profit constraint or which of various conflicting objectives was supposed to dominate.

Following a different line of attack, Herbert Simon (1959, 1962), Richard Cyert and James March (1963), and others generally associated with the Carnegie School rejected maximizing behavior and focused on the process of decision making within the firm. Key notions of this approach included satisficing, multiple goals, organizational slack, resistance to change, and other "behavioral" characteristics (Richard Day, 1964). Although these notions did not provide a unified framework capable of dis-

<sup>3</sup>For a stimulating discussion of some aspects of neoclassical theory and its development, see Melvin Reder (1982).

<sup>4</sup>A detailed discussion of the various efficiency conditions and of some of the difficulties introduced by such things as increasing returns to scale may be found in Francis Bator (1957, 1958) as well as in most advanced texts in microeconomics (for example, see Jack Hirshleifer, 1980).

placing neoclassical theory, they helped to identify some limitations of that theory and to stimulate its revision. Indeed, Williamson (1963, 1964) took several steps toward reconciling various strands of the Carnegie School with orthodox theory by examining a utility-maximization framework in which managers explicitly sought discretionary emoluments.<sup>5</sup> Leibenstein's development of X-efficiency (for example, selective rationality, effort discretion) owes much to the Carnegie School (see his 1979a essay).

A third line of revision sought to generalize the neoclassical framework. Thus, Gary Becker (1957) showed that nonpecuniary sources of utility could be taken explicitly into account in predicting the employment choices of income earners. At about the same time, Alchian (1959, 1961) began to explore the economic consequences of alternative structures of property rights. Building on these advances, Alchian and Kessel (1962) then sought to provide a rigorous theoretical explanation for John Hicks' (1935) dictum that monopolists are less likely than competitive enterprises to maximize profits. Their analysis began with the observation that monopolies are creatures of the state and thus subject to an explicit or implicit profit constraint. This constraint weakens owners' rights to the wealth of the firm, lowering the cost of nonpecuniary sources of utility and yielding an increase in their consumption. In conjunction with work on transaction and information costs by Coase (1960), George Stigler (1961), and others, these contributions opened the way for a major overhaul of neoclassical theory.

## II. Generalized Theory of Choice

At least some of the shortcomings of neoclassical theory arise because its antecedent conditions do not always hold. Thus, in a particular domain, the property rights to resources may not be fully allocated or privately held, and transaction costs may be positive. Failure to take such deviations into

account would yield at least some implications not supported by experience.

Neoclassical theory, however, can be generalized to eliminate some of these limitations. A major step is to end the dichotomy between the theory of consumer choice and the theory of the firm by extending the utility-maximization hypothesis to all individual choices, including those made by business managers and government employees. Another step is to broaden the concept of the limits on individual choices to include institutional constraints (the system of property rights) as well as more of the constraints (for example, including transaction and adjustment costs) imposed by nature and the state of the arts.

Private property means that an individual's rights to the use of the resources he owns are exclusive and voluntarily transferable. If transaction costs are zero, then these rights will be fully defined, fully allocated, and fully enforced. Moreover, they will be reallocated to their highest-valued use regardless of their initial assignment (Coase, 1960).<sup>6</sup>

The existence of positive transaction costs, however, introduces a new constraint and yields new efficient solutions. For example, it implies that some monopolists will choose (efficiently) not to engage in perfect price discrimination. It also implies the (efficient) queuing of resources (for example, commodity inventories, resource unemployment) and of consumers. Further, it implies that some rights to resources will not be fully assigned (for example, some fisheries will be owned in common), fully enforced (some theft will be allowed), or priced (for example, parking spaces in some privately owned shopping centers will be assigned on a first-come, first-served basis), thereby reducing an individual's incentive to take fully into account all the harms and benefits flowing from his decisions.

Moreover, the organization of economic activity within a society need not be based on private property. Other structures of property rights may be chosen, either be-

<sup>5</sup>For a recent summary of managerial theories of the firm, see Marris and Dennis Mueller (1980). Simon still champions satisficing behavior (1979).

<sup>6</sup>Of course, the initial assignment of rights to the use of resources affects the distribution of wealth and thus may indirectly affect the final allocation of resources.

cause—all other things being the same—they yield more utility or, much more likely, because they work to the advantage of specific groups which enjoy a comparative advantage in the use of political power. Be that as it may, different systems of property rights present decision makers with different structures of incentives, resulting in different alignments of resources and different input-output mixes (Alchian, 1965a, 1967).<sup>7</sup>

Focusing on business choices, some of the most recent and significant applications of the property rights/transaction costs generalization yield testable hypotheses regarding the clustering of resource rights, including the choice and evolution of alternative forms of business enterprise (see Alchian and Demsetz, 1972; Alchian, 1979; Williamson, 1975, 1979; Charles Goetz and Robert Scott, 1981). From this perspective, business enterprises develop to solve the shirking-information problem of team production by lowering the cost of monitoring exchanges (including effort) and of directing the allocation of jointly cooperating units (Alchian and Demsetz, 1972). To provide more effective monitoring, the owners of the assets most specialized to the firm (those assets whose value in the next-best use is considerably below their value in the present coalition) typically are residual claimants (Klein et al.).

Following this line of reasoning, some vertical integration may be explained as a lower-cost alternative to enforcing contracts designed to inhibit opportunistic behavior among firms in the same production-distribution chain.

Employees, of course, may also embody firm-specific assets. This event increases the possibilities for opportunistic behavior within the enterprise, fostering the development of institutional and contractual arrangements to control it (see John Cable and Felix FitzRoy, 1980), and generating its own (efficient) equilibrium solution.

Shirking is further inhibited by competition among actual and prospective members of the team (including managers) as well as by the market for control of the enterprise (see Henry Manne, 1965; Williamson, 1970; Eugene Fama, 1980; my 1973 article).

As the size and complexity of a team increase, monitoring costs eventually may exceed the gains from joint production, inhibiting the size of business enterprises (see Williamson, 1967). Monitoring costs also help to explain heterogeneity of firm sizes within each industry (see Walter Oi, 1981) as well as the choice of specific types of business organization. Thus, if monitoring costs are high relative to benefits, and if team production still yields more output than separate operation, then profit-sharing arrangements will evolve to discourage shirking. Examples are partnerships in professional and intellectual work as well as share contracts in agriculture and mining (see Cheung, 1969; William Hallagan, 1978).

When team size can be relatively large, the problem of raising large sums of equity capital encouraged the development of the modern corporation with transferable shares (see Robert Ekelund and Robert Tollison, 1980) as a device for economizing on transaction costs (see Williamson, 1981). Shareholders own the specialized assets and bear the value consequences of exogenous events as well as of the decisions made within the firm. Other individuals typically specialize in deciding how resources are to be used within the corporation, acting as agents for stockholders. The principal-agent relationship is currently being explored in more detail (Michael Jensen and William Meckling, 1976, 1979; Fama, 1980).

Even under a system of fully allocated private property rights, the existence of positive transaction costs implies that there will be some shirking and other deviations from

<sup>7</sup>For example, under common ownership, individuals typically lack exclusive, transferable rights to the use of resources. Relative to a system of private property rights, this implies that individuals will invest less in the commonly held resource and will prefer shorter- to longer-lived investments. It also implies increased entry to capture rents, a lower marginal product of capital, and earlier exhaustion of the resources held in common (for example, Steven Cheung, 1970; Richard Agnello and Lawrence Donnelley, 1975). To mitigate these consequences, the analysis suggests that common ownership is characterized by institutional arrangements that act as surrogates for at least some of the constraints imposed by the market under private ownership rules. Thus, such things as hunting and fishing seasons, limits on catches, specification of admissible harvesting techniques, and similar controls may be expected to occur more frequently under common rather than private ownership.

the efficiency conditions of neoclassical theory. If, in addition, private property rights are attenuated (for example, mutual ownership, government regulation) or replaced by some other institutional arrangement (government ownership, worker management), recent work suggests that deviations from neoclassical conditions will be even more significant (see Kenneth Clarkson and Donald Martin, 1980).

For example, the crucial difference between private and political (publicly owned) firms is that ownership in the latter effectively is nontransferable. This situation rules out specialization in their ownership, inhibiting the capitalization of future value consequences into current transfer prices and reducing the incentive of those who bear such consequences to monitor managerial behavior (see my 1969 article). As a result, managers of political firms typically have greater opportunity for discretionary behavior—as judged by market standards—than do managers of privately owned firms. This implies that private firms are more likely than comparable government-owned firms to introduce cost-reducing innovations, to adopt cost-minimizing input combinations, to incur lower operating costs, to produce a greater variety of output, to use less capital-intensive production techniques, and to incur lower production costs. The evidence supports these and other implications of the analysis (see my 1980 article).

Taking transaction costs and the structure of property rights into account is thus beginning to yield insights not only into why firms exist, but also into the choice of particular kinds of business organizations. Further light on these and other issues is also being cast by work on the evolution of different systems of property rights (see Demsetz, 1967; Gary Libecap, 1978; John Umbeck 1981).

### III. Economic Efficiency

Neoclassical conditions yield an equilibrium which, as noted earlier, is Pareto efficient. As a first departure from the neoclassical world, suppose that transaction costs are positive and rise at the margin. In the case of monopoly, this will inhibit perfect

price discrimination (as well as such things as consumers' cooperatives). In equilibrium, the price of the monopolized commodity will exceed its marginal cost, and Pareto efficiency will no longer hold. The nondiscriminating monopolist will still produce somewhere along its least-cost expansion path and its output configuration will still be on its production possibility curve. Relative to both the competitive solution (assuming that the marginal cost of the monopolist somehow is identical to the competitive supply curve) and to the perfectly discriminating solution (both solutions entail the same output and, at the margin, the same price), the output will be smaller and not all mutually beneficial trade will be exhausted. The resulting loss in producers' and consumers' surplus is usually described as the efficiency or welfare loss from monopoly.

The loss of efficiency, however, is relative to an environment of zero transaction costs. If the costs of enforcing perfect price discrimination are greater than the surplus the monopolist could obtain (in the limit, all of the consumers' and producers' surplus), then the nondiscriminating monopoly solution is efficient. The efficiency loss at issue is analogous to the welfare loss associated with the positive price of any resource (see Demsetz, 1969). For example, the existence of transportation costs means that the delivered price to consumers away from the plant is greater than the marginal cost of the commodity prior to shipment. The higher supply curve entails a welfare loss relative to a world of zero transportation costs.

More generally, suppose that transaction costs are positive and rise at the margin. It is then profitable for the firm to allocate some resources to acquiring information and to drafting and enforcing contracts. Typically, however, the equilibrium solution will not entail either the acquisition of full information, or the drafting and enforcement of all conceivable contractual provisions. As a result, some resources will be used within the firm to produce noncontractual, job-related sources of utility (including shirking). Among other things, this implies that the firm will be operating off its neoclassical least-cost expansion path and will be inside its neoclas-

sical production possibility curve. Nevertheless, if transaction costs are positive as stipulated, the solution must be efficient.<sup>8</sup>

To carry the analysis further, consider the consequences of alternative systems of property rights, holding transaction costs constant. Under reasonable conditions, shifting from private to common ownership would alter the structure of incentives (for example, discourage individual investments in commonly held property) in such a way that production would now be inside the neoclassical production possibility curve. If resources are held in common, however, the "interior" solution again is efficient. Indeed, if common ownership were to yield more of other outputs (for example, the utility of the system) not captured by the neoclassical solution, common ownership might well be preferred by all members of the community.

As the preceding comments suggest, efficiency is being defined as constraint maximization. Efficiency conditions are seen as the properties of a determinate (equilibrium) solution implied by a given theoretical construct. On this view, a system's solutions are always efficient if they meet the constraints that characterize it.<sup>9</sup>

#### IV. The X-Efficiency Construct

In his original 1966 paper, Leibenstein began by noting that the relatively small

welfare losses associated with monopoly (Arnold Harberger, 1959; David Schwartzman, 1960) and with tariffs (Johnson, 1958; Scitovsky, 1958) were based on the assumption that firms were minimizing costs at the level of output at which they were producing.<sup>10</sup> Having suggested that failure to minimize costs (X-inefficiency) was common and could yield substantial welfare losses, Leibenstein then presented findings from several case studies of U.S. and foreign firms which seemed to support his argument.

Leibenstein's criticism was well taken, although his statement of the case for X-inefficiency was fragmentary. For example, he did not take into account property rights structures (for example, whether the firms examined were privately or government owned), and the reasons he offered for the existence of X-inefficiency appeared to be straightforward implications of positive transaction costs.<sup>11</sup> Leibenstein, however, has honed his argument a good deal since 1966. Accordingly, it seems best to focus on more recent statements of his position.

In a recent summary of X-efficiency, Leibenstein opened with the statement that

The view behind this paper is that although neoclassical (NC) micro theory works some of the time, there are areas of experience to which it is not applicable. As a consequence, it is desirable to develop models which are more general than the NC framework, which fit economic realities, and into which the NC framework fits as a special case.

[1978c, p. 328]

Granting the limitations of traditional neoclassical theory, the desirability of developing more general models at best is debatable. Such models would not be particularly

<sup>8</sup>William Comanor and Leibenstein (1969) argued that estimates of the welfare loss from monopoly have failed to include the loss due to the less "efficient" use of inputs by the monopolist. Michael Crew and Charles Rowley (1971) have elaborated the point, suggesting that it offsets second-best considerations and thus strengthens the case for antitrust. Ross Parish and Yew-Kwang Ng (1972), however, using W. M. Corden's (1970) analysis as a springboard, showed that this additional loss vanishes in the limit if the costs of increasing "efficiency" are taken into account, if the leisure produced within the firm is considered to be a good, and if decision makers bear the full cost of substituting leisure for money income. Similarly, Harry Johnson (1970) argued that the concept of efficiency must include the entrepreneur's leisure (and, presumably, all other sources of utility produced within the firm).

<sup>9</sup>In his critique of Kenneth Arrow (1962), Demsetz (1969) argued that it is inappropriate to judge the efficiency of real world institutions by comparing them to some ideal norm.

<sup>10</sup>Estimate of welfare losses varied from 1/10 to 7/10 of 1 percent of GNP in the case of monopoly and from near zero to a maximum of 1 percent in the case of tariffs (see Leibenstein, 1966, p. 393).

<sup>11</sup>Leibenstein offered three reasons for X-inefficiency: "(a) contracts for labor are incomplete, (b) the production function is not completely specified or known, and (c) not all inputs are marketed or, if marketed, are not available on equal terms to all buyers" (1966, p. 412).

useful unless there were rules specifying *ex ante* which model would be applicable to a particular set of circumstances, a situation which suggests the existence of a more general underlying theory. Accordingly, it would be more desirable (useful) to seek the latter. This does not deny that moves toward a more general theory might well rest on the rubbles of discarded *ad hoc* models which temporarily filled gaps in knowledge.

The central difficulty with X-efficiency is that it focuses on preference relations that are not observable and thus fail to yield testable hypotheses. Consider the basic postulates and related variables of X-efficiency theory that Leibenstein lists and contrasts with those of neoclassical theory (see his Table 1, 1978c, p. 329). The axiom of selective rationality asserts that individuals choose the extent to which they deviate from maximizing behavior (1979b, p. 485), with the degree of deviation determined by the personality of the individual and the economic context (1979b, p. 485). Similarly, inert areas are determined by the inertial cost of moving, which "...depends on an individual's personality" (1978c, p. 329). The postulate of incomplete contracts implies the existence of effort discretion (1979b, p. 485), the amount of effort being determined by constraint concern (1979b, p. 487) which in turn is determined by personality and the economic context (1978c, p. 328). The divergence of interests between principal and agent is due at least in part to inertial costs (1972, p. 327), and thus also depends on personality.

To use X-efficiency for predictive purposes, it is necessary, among other things, to attach meaning to the term personality. According to Leibenstein, "Personality is defined in terms of (b) a taste for responsiveness to opportunities and constraints within certain standard of behavior and (c) a simultaneous taste for 'irresponsible' or 'unconstrained' behavior" (1979a, p. 485). These tastes, however, are not observable. It follows that selective rationality, inert areas, discretionary effort, and other axioms and related variables of X-efficiency used by Leibenstein from time to time are not operational beyond yielding the general conclusion, built directly into the axioms, that firms do not produce on their neoclassical least-cost

expansion paths. Indeed, the axioms of X-efficiency seem to imply that the wealth-maximization hypothesis never holds, a position stronger than seems warranted by Leibenstein's evidence.

In contrast, generalized neoclassical theory focuses on changes in constraints, which are potentially observable. To derive and test implications of the theory, it is sufficient to identify the commodities that enter an individual's utility function—something that X-efficiency also must do—and then indicate how observed changes in constraints affect the relative cost of the commodities to the individual making the choice.

Leibenstein's collection of postulates and related variables of X-efficiency appears to be a combination of some of the axioms and some of the implications of generalized neoclassical theory. For example, a key component of X-efficiency is that the individual, rather than the household and the firm, is the basic unit of analysis. Thus, "...the most important distinction between the X-efficiency theory and the neoclassical model of the firm is the fact that in X-efficiency theory the individual is the basic atomic unit" (1979a, p. 486), rather than the firm and the household.

The application of the utility-maximization hypothesis to all choices, of course, rests on the individual as the basic unit of analysis. Indeed, it explicitly ends the dichotomy between the theory of consumer choice and the theory of the firm, thereby eliminating the most important distinguishing characteristic of X-efficiency.

A second postulate or component of X-efficiency is that an individual's attentiveness to opportunities for gains and to constraints that can impose losses depends on his personality and on the economic context. That is, there is selective rationality rather than maximizing (or minimizing) behavior.<sup>12</sup>

<sup>12</sup>According to Leibenstein, "A maximizer would take advantage of all opportunities for *gain*, and attend fully to *all* constraints which, if not attended to, would impose a *loss*" (1978c, p. 328). This view of maximizing behavior is quite similar to Thorstein Veblen's (1919) caricature of economic man:

The hedonistic conception of man is that of a lightning calculator of pleasures and pains, who oscillates like a homogeneous globule of desire of happiness under the

A utility-maximizing individual, however, does have the incentive to respond selectively. Faced by positive transaction and adjustment costs, an individual will respond only to those events that are sufficiently large or long lasting to justify a change. Thus, holding other things constant, an increase in transaction or adjustment costs implies an increase in the threshold that must be overcome before an individual chooses to respond. Moreover, holding other things constant, the attenuation of an individual's property rights reduces the gains that he will be able to capture, and thus reduces the incentive to respond to a given change in constraints. Accordingly, differences in transaction and adjustment costs and in the structure of property rights affect systematically (and predictably) the threshold and the extent to which individuals respond to a change in constraints. If this is what Leibenstein means by selective rationality, then it is an implication of the utility-maximization hypothesis.<sup>13</sup>

Perhaps the most telling criticism of selective rationality as an axiom of X-efficiency is provided by Leibenstein himself. Thus,

Using the ideas developed in the previous paragraphs I develop what may be viewed as the basic proposition of X-efficiency theory. Individuals' contracts and effort choices impose costs on the

firm, and the firm attempts to obtain "budgets" from the "environment" to at least cover such costs. It is assumed that individuals are so motivated that they prefer less confining constraints to more confining ones. This translates to a desire for greater rather than smaller individual budgets. [1978c, p. 330]

That is, individuals prefer more to less: they maximize utility. In order to derive the basic proposition of X-efficiency, Leibenstein here appears to have jettisoned the axiom of selective rationality and adopted the utility maximization postulate.

The other components of X-efficiency theory listed by Leibenstein<sup>14</sup> also appear to be straightforward implications of the generalized neoclassical theory. Thus, a third postulate asserts the existence of inert areas because an individual "...will not necessarily move to a superior position in the standard utility sense because of the inertial costs of moving" (1978c, p. 329).

The same arguments presented in the case of selective rationality apply. That is, the existence of positive transaction and adjustment costs and the attenuation of private rights to the use of resources imply that an individual will respond only if the present-value benefits are greater than the present-value costs associated with the change.

Leibenstein seems to recognize the point. In a reply to a comment by K. J. Blois (1972) regarding conflict between managers and shareholders, Leibenstein notes that the relevant costs of overcoming inertia are "...the *utility costs* to the owners of replacing the managers" (1972, p. 327), which is certainly an important part of the answer.

Fourth, contracts are incomplete, and fifth, effort is a discretionary variable. Both phenomena are implied by the costs of monitoring and of enforcing contracts as well as by differences in incentive associated with different structures of property rights. The shirking problem is a central point of the Alchian and Demsetz paper (1972) and of

impulse of stimuli that shift him about the area, but leave him intact. He has neither antecedent nor consequent. He is an isolated, definitive human datum, in stable equilibrium except for the buffets of impinging forces that displace him in one direction or another. Self-imposed in elemental space, he spins symmetrically about his own spiritual axis until the parallelogram of forces bears down upon him, whereupon he follows the line of the resultant. When the force of the impact is spent, he comes to rest, a self-contained globule of desire as before. [pp. 73-74]

Both views seem to be equally out of touch with current theoretical and empirical work.

<sup>13</sup>As Leibenstein argues, individuals no doubt differ regarding the way they feel they ought to behave and the way they would like to behave. They also differ in a variety of other characteristics (psychological, intellectual, physical) too numerous to list, and these differences presumably affect the extent and the intensity with which individuals respond to various changes in circumstances. What is at issue, therefore, is the usefulness of the theory used to choose and organize the variables that matter.

<sup>14</sup>The number of postulates and basic variables of X-inefficiency theory is somewhat flexible. Leibenstein listed three in 1966, nine in March 1978 (Table 1, p. 204), and six in May 1978 (Table 1, p. 329).

the modern literature on the behavior of business enterprises (Williamson 1975, 1979; Klein 1980). Indeed, it is a crucial consideration in explaining why firms exist and the choice of organizational form.

Finally, Leibenstein argues that there is a difference of interests among principals and agents rather than the identity presumed by neoclassical theory. Again, positive transaction costs and deviations from private property rights imply such a differential. As noted earlier, the literature exploring the economics of principal-agent relationship in the context of a generalized neoclassical theory is growing and fruitful.

On purely methodological grounds, the X-efficiency construct may be dismissed by the rule of Occam's razor as a step backwards in the development of economic theory. Whereas generalized neoclassical theory offers fewer, more general axioms yielding a richer, more powerful set of testable implications, X-efficiency offers instead a deductive system with more axioms that seek to be more descriptive yet yield fewer, less clearly specified implications. Indeed, it is difficult to escape the conclusion that Leibenstein is merely trying to offer a more descriptively "realistic" set of axioms. This is a methodological route that, for good reasons (see R. Schlegel, 1967), may be expected to lead to a dead end (see Milton Friedman, 1953; Nagel, 1963; Lawrence Boland, 1979).

Given the pick-and-choose nature of X-efficiency, it is possible that some of the hypotheses which Leibenstein claims to derive from it cannot be derived from generalized neoclassical theory.<sup>15</sup> Accordingly, the main propositions of X-efficiency are examined next.

#### V. The Main Propositions of X-Efficiency

The central proposition of X-efficiency seems to be that not all firms minimize costs; that is, not all firms produce on the outer bounds of their production possibility surfaces. As a corollary, "...firms do not always introduce technical changes when

available and profitable" (Leibenstein, 1969, p. 600).

In the absence of any indication when firms may be expected to behave this way, however, these assertions are not very useful. A recent exchange between Stigler (1976) and Leibenstein (1978b) offers an excellent point of departure for exploring what really is at issue. Stigler noted that Leibenstein (1966) has ascribed X-inefficiency to motivational deficiencies and to inefficient markets for knowledge. Regarding the motivational issue, Stigler argued that if transaction costs are positive (presumably the rights to the use of resources are privately held), then contracts are not fully specified and enforced. As a result, some resources are used to enforce contracts and other resources, due to the incomplete enforcement of contracts, are used to produce other sources of utility (for example, more leisure and less guns and butter). Regarding the choice of production technique, Stigler argued that all firms operate on their production possibility curves (PPC). Different firms, however, may be on different PPCs due to differences in entrepreneurial capacity or investment in knowledge. Stigler concluded that inefficiency can arise only *ex post* because of *ex ante* estimating errors.<sup>16</sup>

Leibenstein's reply focused on the choice of technique and consisted largely of a restatement of his case for X-inefficiency. He added, however, that Stigler presented no reason for the value of the additional leisure produced within the firm "...to be in any sense equal to or greater than the reduced value of the product" (1978b, p. 209). Leibenstein concluded that if the money value of the compensating increase in job-related utility is less than the product foregone, then there is X-inefficiency.

From Stigler's perspective, the point is irrelevant. Subject to the constraints applicable to a particular situation, economic forces yield the equilibrium solution. In equilibrium, all the anticipated costs (search,

<sup>15</sup>In such an event, the implications presumably could not be derived from the axioms of X-efficiency either.

<sup>16</sup>According to Stigler, inefficiency could arise in the absence of uncertainty because of a failure of economic agents to maximize, a view which he considers to be a methodological leap into the unknown.

transaction, adjustment, etc.) of realigning the resources are necessarily greater than the anticipated value of the additional benefits that could be obtained. Thus, the equilibrium solution associated with a given set of constraints is efficient. As a corollary, the distinction Leibenstein draws between allocative efficiency and X-efficiency is spurious. The only issues are allocative.

The definitional debate thus masks a more fundamental disagreement regarding the role of theory. Stigler's definition of efficiency reflects a concern with economics as an engine of analysis for predicting choices. Leibenstein, on the other hand, seems to use efficiency as a normative concept to describe some ideal solution. Deviations from that ideal are then defined as inefficient without regard to real world alternatives.

In the real world, however, different institutional arrangements typically confront decision makers with different structures of property rights as well as with different structures of transaction costs. As a result, different economic systems may be expected to yield different configurations of inputs and outputs, as well as of wealth. In turn, such differences presumably affect the choice of institutions.

Within the X-efficiency framework, competition plays a powerful role in forcing the higher-cost producers to reduce costs or to exit. Leibenstein concludes that "... there is a lower cost level for the average firms under competition than under monopoly" (1973a, p. 776), and reports approvingly evidence to that effect.<sup>17</sup>

Generalized neoclassical theory yields similar results. As an initial approximation, suppose that transaction costs are positive and rise at the margin. Moreover, suppose that all business enterprises, whether monopolistic or competitive, are single proprietorships, are exempt from explicit or implicit govern-

ment regulation, are price takers in the input markets, and are in long-run equilibrium.

Under pure wealth-maximizing conditions, both types of enterprises will incur some costs to acquire information about prices and other relevant variables as well as to monitor and enforce contracts. All other things being the same, however, the monopolist (a price searcher) will face higher costs in seeking to locate the demand curve for its output than a competitive firm (a price taker in the limit). As a result of acquiring demand information, therefore, monopolists typically will incur higher production costs. Moreover, because demand information is costlier for the monopolists, they will acquire relatively less of it. As a result, monopolists' output-price combinations will exhibit greater variance with respect to ideal neoclassical conditions.

In a world of uncertainty and positive transaction costs, the existence of more competitors implies lower production costs and output-price combinations closer to those predicted by neoclassical theory. These results follow simply because in a more competitive environment there are more firms, including potential entrants, searching for the least-cost combination and for the most profitable output-price configuration.

Broadening the analytical framework to allow for utility-maximizing behavior, single proprietors of both monopolistic and competitive firms may be expected to use some of their wealth to acquire job-related, non-pecuniary sources of utility (say, leisure). In a more competitive environment, however, with more firms searching for lower-cost alternatives and more profitable output-price combinations, the opportunity cost of non-pecuniary sources of utility will be higher, and less will be acquired. That is, single proprietors of competitive firms will have less opportunity to indulge their tastes, and production costs of the measured output will be lower.

Finally, consider differences in property rights, including government regulation. Under reasonable conditions, the owners of a wealth-maximizing monopoly subject to a profit constraint have the incentive to adopt a ratio of capital to labor greater than that

<sup>17</sup>Schwartzman (1973) has questioned Leibenstein's argument that competitive firms are necessarily more efficient than monopolists. Mark Crain and Asghar Zardkoobi (1980) have made the same point, while noting the substantial costs of activities designed to capture and maintain monopoly rents (Gordon Tullock, 1967; Richard Posner, 1975).

which would have minimized production costs at the output level chosen (see Harvey Averch and Leland Johnson, 1962). Stricter regulation (for example, decreasing the lag in the adjustment of the actual to the allowed rate) reduces the incentive to adopt cost-reducing and demand-increasing innovations (Baumol and Klevorick, 1970). An active profit constraint, moreover, also weakens owners' property rights, reducing their incentive to monitor, and thereby increasing managers' and employees' opportunities for discretionary behavior.<sup>18</sup> Several implications of this analysis have been tested, and the evidence supports them (Alfred Nicols, 1967; Franklin Edwards, 1977).

As Mark Crain and Asghar Zardkoohi (1980) point out, Leibenstein's statement of X-efficiency implies that the ownership of the firm does not matter. Accordingly, X-efficiency would be unable to explain observed differences in behavior among such organizational forms as privately owned corporations, mutuals, cooperatives, and worker-managed firms. Indeed, it would not even distinguish between privately owned and publicly owned firms.

Within the generalized neoclassical framework, competition inhibits shirking and thus, among other things, yields lower costs. Competition from other firms in the industry and from prospective entrants provides a check on performance, encourages the evolution of internal control devices, and ultimately eliminates higher-cost firms unable or unwilling to adjust. Competition in the capital market acts to transfer ownership (control) of the enterprise to those better able to use it, while competition for managerial and team positions from candidates both within and without the enterprise discourages shirking by employees. Transaction costs and the structure of property rights affect both the extent and the form of shirking.

## VI. The Evidence of X-Efficiency

The X-efficiency construct does not provide a systematic framework for predicting

when and how firms will fail to minimize costs. Accordingly, it is not surprising that Leibenstein's approach to "testing" the X-efficiency construct consists of noting observation statements in the literature to the effect that firms do not minimize costs.

In Leibenstein's original article (1966), the evidence he offered included impressionistic findings (interviews) by Erik Lundberg (as reported by Goran Ohlin, 1962), Frederick Harbison (1956), and Neil Chamberlin (1962) that some firms could have produced the same output at a lower cost. It also included Peter Kilby's (1962) summary of the results of several International Labor Organization (ILO) productivity missions, indicating that simple reorganization of production frequently yielded substantial (more than 25 percent) cost reductions. Interestingly, after the missions left, some if not all firms returned to previous methods and productivities. Leibenstein also noted a study by Laszlo Rostas (1964), who examined a variety of U.S. and U.K. industries and found differences in productivity that he could not attribute to differences in equipment, as well as studies by J. P. Davison et al. (1958), by the ILO (1951), and others showing that changes in pay structures affect productivity per man.

Leibenstein (1966) then turned to the issue of technical innovation. Here he reported findings by C. F. Carter and B. R. Williams (1958) that a high proportion of investment in innovation is defensive, by W. E. G. Salter (1960) and others that innovations are adopted slowly, and by John Johnston (1963) that in Great Britain the average rate of return on consulting fees was 200 percent. He also noted micro- and macroeconomic findings suggesting that motivational factors account for much of the growth of output (1966, pp. 403-05).

In a book-length treatment of X-efficiency, Leibenstein (1976) reiterated the evidence just described and added the following studies whose authors typically adduced X-efficiency to explain some of the results observed. Thus, Joel Bergsman (1974) argued that firms protected from foreign competition relax cost-reducing efforts and estimated that the higher production costs due to such X-inefficiency accounted for 0.4 to 6

<sup>18</sup>Regulation, whether explicit or implicit, yields a new, efficient equilibrium solution (Alchian and Kessel, 1962; Ross Parish, 1970).

percent of the gross national product of six developing countries. Along similar lines, Walter Primeaux (1977) reported that electric utilities which faced competition had lower costs than those which had a monopoly.

Further, Tsung-Yuen Shen (1973) used a sample of 4,000 manufacturing plants in Massachusetts to explore the relationship between output, capital, and labor inputs. After allowing for scale and technology diffusion, he found that the pattern of residual variation was best explained by efficiency "...or with less ambiguity and more style, X-efficiency" (p. 273). Best-practice plants enjoyed both technology and efficiency, whereas other plants enjoyed only technology. William Shepherd (1972), in reporting the results of a study of 231 large U.S. firms, noted that "Size carries a negative coefficient, perhaps owing to X-inefficiency of large absolute size" (p. 35), and that "If actual profits have reflected X-inefficiency (the gap between actual and attainable profits), then market share-reduction could also yield rises in X-efficiency" (p. 35).

Turning to other sources of X-inefficiency, Leibenstein reported findings by John Shelton (1967) that the same franchises were more profitable when operated by owner-managers than by experienced managers employed by the parent company. After also noting work by H. K. Radice (1971) and others showing that "...where the ownership component of management is higher, profit rates are higher" (1976, p. 44), he opined that the relationship between ownership and efficiency would be even stronger.

Elsewhere, Leibenstein (1973a) cites a paper by Michael Crew and Charles Rowley (1971) in which they suggested that Williamson's (1964) evidence of discretionary behavior provides an example of "overhead" X-inefficiency. He also takes as evidence of X-inefficiency (1977a, p. 316) the finding by Kenneth Shapiro and Jürgen Müller (1977) that cotton farmers in Tanzania do not make full use of the information they have.

As Leibenstein recognizes, the kinds of evidence he offers in support of X-efficiency "...may invite the charge of casual empiricism" (1973a, p. 777). Rather than assess the validity of individual studies, however, for

present purposes it is more useful to stipulate their findings and to explore instead whether such evidence could be inconsistent with generalized neoclassical theory.

Several of the studies noted by Leibenstein report the consequences associated with alternative constraints, and these results are easily explained by generalized neoclassical theory. Thus, for example, high monitoring costs encourage profit-sharing arrangements (Alchian and Demsetz, 1972). This explains the findings by Shelton (1967) that the same franchises are more profitable when operated by owner-managers than by employees of the parent company and by Radice (1971) that increasing the ownership component of management yields higher profit rates.<sup>19</sup> On the other hand, low cost of metering physical output encourages tying workers' compensation to their output, explaining observations by Davison et al. (1958) and others (ILO, 1951) that under some circumstances, piece rates yield an increase in productivity. Similarly, transaction costs and the weakening of owners' property rights associated with explicit or implicit regulation could account for the higher costs reported by Primeaux (1977) for electric utilities which have a monopoly and by Bergsman (1974) for firms protected from foreign competition, and for the negative correlation between size (and market share) and profits observed by Shepherd (1972).<sup>20</sup> Along similar lines, Kilby's (1962) finding that increases in productivity under the aegis of ILO demonstration teams dropped back to old levels after the teams' departure suggests that constraints were changed only during the demonstration period, so that afterwards firms drifted back to their old equilibrium positions (Clarkson, 1980). Williamson's work (1964) on discretionary behavior, of course, was one of the

<sup>19</sup>For a discussion of the nature of the employment contract and its effect on effort, see also E. Odgers Olsen (1976) and Joseph Stiglitz (1975).

<sup>20</sup>As noted earlier, monitoring costs presumably impose a limit on firm size (see Williamson, 1967; Alchian and Demsetz, 1972). Larger firms and larger market shares also attract government attention—and explicit or implicit controls—that inhibit growth and increase opportunities for discretionary behavior.

early contributions to the development of the generalized neoclassical theory.

The other studies cited by Leibenstein, however, merely assert or document deviations from idealized neoclassical theory. Thus, Ohlin (1962), Harbison (1956), Chamberlin (1962), Rostas (1964), and Shapiro and Müller (1977) simply report that output could have been produced at a lower cost, an observation implied by a variety of alternative property rights/transaction costs conditions. Similarly, Carter and Williams (1958), Salter (1960), Johnston (1963), and Shen (1973) report differences in the adoption of innovations but do not provide sufficient information regarding possible differences in property rights, transaction costs, or other variables that could have accounted for the behavior observed. None of these findings are inconsistent with the generalized neoclassical theory presented in this paper.<sup>21</sup>

The empirical studies used by Leibenstein to support the existence of X-efficiency, with a few notable exceptions such as Shelton (1967) and Primeaux (1977), do not take differences in transaction costs and in property rights into account. Typically, instead, the actual performance of a sample of firms with respect to such variables as costs, profits, or the introduction of innovations is compared to the performance expected under ideal neoclassical conditions. If the ideal equilibrium solution does not hold, the failure is then ascribed to unknown variables and labeled X-inefficiency (Shen, 1973, p. 273). The information provided in these studies, therefore, typically is insufficient to determine why the firms behaved as they did. As a corollary, it is also insufficient to evaluate the generalized neoclassical theory, thereby failing to deny it.

<sup>21</sup>The transaction costs/property rights approach presents quite a general framework for dealing with the phenomena at issue. For example, it is more general than those proposed by Ken Jameson (1972), who sought to explain observed instances of X-efficiency by focusing on dynamic considerations, and by Crew et al. (1971), who sought to salvage the profit-maximization theory of the firm by taking into account policing expenditures aimed at curbing X-inefficiency. It also offers a much broader concept of competition than Paul McNulty (1967) envisioned.

## VII. Conclusion

Economists have recognized for some time that neoclassical economic theory has serious limitations. Since the early 1960's, a great deal of work has sought to revise the theory and to extend its range of application by focusing on the individual as the basic unit of analysis and by taking into account the constraints imposed by positive transaction and adjustment costs as well as by the structure of property rights. A substantial and growing body of theoretical and empirical research suggests that this generalization is fruitful.

Leibenstein has ignored this literature and focused his criticism on the failure of "traditional" or "textbook" neoclassical theory to allow for organizational and frictional elements. The X-efficiency construct which he offers as an alternative, however, seems to be an amalgam of some of the axioms and some of the implications of generalized neoclassical theory. To the extent that X-efficiency has any predictive content, it is already encompassed by the more general neoclassical theory which yields a broader and richer set of implications.<sup>22</sup> In practice, Leibenstein and others seem to use X-inefficiency simply as a catch phrase to denote *ex post* deviations from idealized neoclassical equilibrium conditions.

<sup>22</sup>Leibenstein himself seems to be undecided between X-efficiency and the utility-maximization hypothesis. For example, he states that:

The basic phenomenon to be explained is that frequently individuals, groups, or firms do not take advantage of opportunities for gain. Broadly put, the central explanatory mechanism is that, while such opportunities would increase utility, the costs of making the change from an existing effort level to the required effort level to take advantage of the gain is greater than the utility of the net gain involved. [1969, p. 602]

This clearly describes utility-maximizing behavior. Taking account of the constraints imposed by differences in property rights and transaction costs would help to identify the incentive structure faced by individual decision makers and would allow prediction of their choices—including why they do not take advantage of what, under a narrower perception of constraints, appear to be opportunities for gain.

## REFERENCES

- Agnello, Richard J. and Donnelley, Lawrence P., "Prices and Property Rights in the Fisheries," *Southern Economic Journal*, October 1975, 42, 253-62.
- Alchian, Armen A., "Private Property and the Relative Cost of Tenure," in Philip D. Bradley, ed., *The Public Stake in Union Power*, Charlottesville: University Press of Virginia, 1959, 350-71.
- \_\_\_\_\_, *Some Economics of Property*, Santa Monica: Rand Corporation, 1961.
- \_\_\_\_\_, (1965a) "The Basis of Some Recent Advances in the Theory of Management of the Firm," *Journal of Industrial Economics*, December 1965, 14, 30-41.
- \_\_\_\_\_, (1965b) "Some Economics of Property Rights," *Il Politico*, December 1965, 30, 816-29.
- \_\_\_\_\_, "How Should Prices Be Set?," *Il Politico*, June 1967, 32, 369-82.
- \_\_\_\_\_, "Property Rights and Organization of Economic Activity," mimeo., November 2, 1979.
- \_\_\_\_\_ and Demsetz, Harold, "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, 62, 777-95.
- \_\_\_\_\_ and Kessel, Reuben A., "Competition, Monopoly, and the Pursuit of Money," in National Bureau of Economic Research, *Aspects of Labor Economics*, Princeton: Princeton University Press, 1962, 157-75.
- Arrow, Kenneth J., "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Princeton: Princeton University Press, 1962, 609-25.
- Averch, Harvey and Johnson, Leland L., "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, December 1962, 52, 1052-69.
- Bator, Francis M., "The Simple Analytics of Welfare Maximization," *American Economic Review*, March 1957, 47, 22-59.
- \_\_\_\_\_, "The Anatomy of Market Failure," *Quarterly Journal of Economics*, August 1958, 72, 351-79.
- Baumol, William J., *Business Behavior, Value and Growth*, New York: Harcourt Brace & World, 1959.
- \_\_\_\_\_, "On the Theory of Expansion of the Firm," *American Economic Review*, December 1962, 52, 1078-87.
- \_\_\_\_\_ and Klevorick, Alvin K., "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion," *Bell Journal of Economics*, August 1970, 1, 162-90.
- Becker, Gary S., *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957.
- Bergsman, Joel, "Commercial Policy, Allocative and 'X-Efficiency'," *Quarterly Journal of Economics*, August 1974, 88, 409-33.
- Blois, K. J., "A Note on X-Efficiency and Profit Maximization," *Quarterly Journal of Economics*, May 1972, 86, 310-12.
- Boland, Laurence A., "A Critique of Friedman's Critics," *Journal of Economic Literature*, June 1979, 17, 503-22.
- Braithwaite, Richard B., *Scientific Explanation*, New York: Harper Torchbooks, 1960.
- Cable, John R. and FitzRoy, Felix R., "Productive Efficiency, Incentives and Employee Participation: Some Preliminary Results from West Germany," *Kyklos*, 1980, 33, 100-21.
- Carter, C. F. and Williams, B. R., *Investment in Innovations*, London: Oxford University Press, 1958.
- Chamberlin, Neil W., *The Firm: Micro-Economic Planning and Action*, New York: McGraw-Hill, 1962.
- Cheung, Steven N. S., "Transaction Costs, Risk Aversion, and the Choice of Contractual Arrangements," *Journal of Law and Economics*, April 1969, 12, 23-42.
- \_\_\_\_\_, "The Structure of a Contract and the Theory of a Non-Exclusive Resource," *Journal of Law and Economics*, April 1970, 13, 49-70.
- Clarkson, Kenneth W., "Managerial Behavior in Nonproprietary Organizations," in his and Donald L. Martin, eds., *The Economics of Nonproprietary Organizations*, Supplement 1 of *Research in Law and Economics*, 1980, 3-26.
- \_\_\_\_\_ and Martin, Donald L., *The Economics of Nonproprietary Organizations*, Supplement 1 of *Research in Law and Economics*, 1980.
- Coase, Ronald H., "The Problem of Social

- Cost," *Journal of Law and Economics*, October 1960, 3, 1-44.
- Comanor, William S. and Leibenstein, Harvey, "Allocative Efficiency, X-Efficiency and the Measurement of Welfare Losses," *Economica*, August 1969, 36, 304-09.
- Corden, W. M., "The Efficiency Effects of Trade and Protection," in I. A. McDougall and R. H. Snape, eds., *Studies in International Economics*, Amsterdam-London: North-Holland, 1970, 1-10.
- Crain, Mark W. and Zardkoobi, Asghar, "X-Inefficiency and Nonpecuniary Rewards in a Rent-Seeking Society: A Neglected Issue in the Property Rights Theory of the Firm," *American Economic Review*, September 1980, 70, 784-92.
- Crew, Michael A., Jones-Lee, M. W., and Rowley, Charles K., "X-Theory Versus Management Discretion Theory," *Southern Economic Journal*, October 1971, 38, 173-84.
- \_\_\_\_\_, and Rowley, Charles K., "On Allocative Efficiency, X-Efficiency and the Measurement of Welfare Loss," *Economica*, May 1971, 38, 199-203.
- Cyert, Richard M. and March, James G., *A Behavioral Theory of the Firm*, Englewood Cliffs: Prentice-Hall, 1963.
- Davison, J. P. et al., *Productivity and Economic Incentives*, London: Allen and Unwin, 1958.
- Day, Richard H., "Review of *A Behavioral Theory of the Firm* by Cyert and March," *Econometrica*, July 1964, 32, 461-65.
- De Alessi, Louis, "Implications of Property Rights for Government Investment Choices," *American Economic Review*, March 1969, 59, 13-24.
- \_\_\_\_\_, "Private Property and Dispersion of Ownership in Large Corporations," *Journal of Finance*, September 1973, 28, 839-51.
- \_\_\_\_\_, "Managerial Tenure under Private and Government Ownership in the Electric Power Industry," *Journal of Political Economy*, May/June 1974, 82, 645-53.
- \_\_\_\_\_, "The Economics of Property Rights: A Review of the Evidence," *Research in Law and Economics*, 1980, 2, 1-47.
- Demsetz, Harold, "Toward a Theory of Property Rights," *American Economic Review Proceedings*, May 1967, 57, 347-59.
- \_\_\_\_\_, "Information and Efficiency," *Journal of Law and Economics*, April 1969, 12, 1-22.
- Edwards, Franklin R., "Managerial Objectives in Regulated Industries: Expense-Preference Behavior in Banking," *Journal of Political Economy*, February 1977, 85, 147-62.
- Ekelund, Robert B. Jr. and Tollison, Robert D., "Mercantilist Origins of the Corporation," *Bell Journal of Economics*, August 1980, 11, 715-20.
- Fama, Eugene F., "Agency Problems and the Theory of the Firm," *Journal of Political Economy*, April 1980, 88, 288-307.
- Friedman, Milton, "The Methodology of Positive Economics," in his *Essays in Positive Economics*, Chicago: University of Chicago Press, 1959, 3-43.
- Furubotn, Eirik G. and Pejovich, Svetozar, "Property Rights and Economic Theory: A Survey of Recent Literature," *Journal of Economic Literature*, December 1972, 10, 1137-62.
- Goetz, Charles J. and Scott, Robert E., "Principles of Relational Contracts," *Virginia Law Review*, September 1981, 67, 1089-150.
- Hallagan, William, "Self-selection by Contracting Choice and the Theory of Sharecropping," *Bell Journal of Economics*, August 1978, 9, 344-54.
- Harberger, Arnold, "Using the Resources at Hand More Effectively," *American Economic Review Proceedings*, May 1959, 49, 134-47.
- Harbison, Frederick, "Entrepreneurial Organization as a Factor in Economic Development," *Quarterly Journal of Economics*, August 1956, 70, 364-79.
- Hicks, John R., "Annual Survey of Economic Theory: The Theory of Monopoly," *Econometrica*, January 1935, 3, 1-20.
- Hirshleifer, Jack, *Price Theory and Applications*, 2d ed., Englewood Cliffs: Prentice-Hall, 1980.
- Jameson, Ken, "Comment on the Theory and Measurement of Dynamic X-Efficiency," *Quarterly Journal of Economics*, May 1972, 86, 313-26.
- Jensen, Michael C. and Meckling, William H., "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure,"

- ture," *Journal of Financial Economics*, October 1976, 3, 305-60.
- \_\_\_\_\_, and \_\_\_\_\_, "Rights and Production Functions: An Application to Labor-Managed Firms and Codetermination," *Journal of Business*, October 1979, 52, 469-506.
- Johnson, Harry G., "The Gains from Freer Trade with Europe: An Estimate," *Manchester School of Economics*, September 1958, 26, 247-55.
- \_\_\_\_\_, "The Efficiency Effects of Trade and Protection: Comment," in I. A. McDougall, and R. H. Snape, eds., *Studies in International Economics*, Amsterdam-London: North-Holland, 1970, 15-17.
- Johnston, John, "The Productivity of Management Consultants," *Journal of the Royal Statistical Society*, 1963, Part 2, 126, 237-49.
- Kafoglis, Milton Z., "Output of the Restrained Firm," *American Economic Review*, September 1969, 59, 583-89.
- Kilby, Peter, "Organization and Productivity in Backward Economies," *Quarterly Journal of Economics*, May 1962, 76, 303-10.
- Klein, Benjamin, "Transaction Cost Determinants of 'Unfair' Contractual Arrangements," *American Economic Review Proceedings*, May 1980, 70, 356-62.
- \_\_\_\_\_, Crawford, Robert G. and Alchian, Armen A., "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, 21, 297-326.
- Leibenstein, Harvey, "Allocative Efficiency vs. 'X-Efficiency'," *American Economic Review*, June 1966, 56, 392-415.
- \_\_\_\_\_, "Organizational or Frictional Equilibria, X-Efficiency and the Rate of Innovation," *Quarterly Journal of Economics*, November 1969, 83, 600-23.
- \_\_\_\_\_, "Comment on the Nature of X-Efficiency," *Quarterly Journal of Economics*, May 1972, 86, 327-31.
- \_\_\_\_\_, (1973a) "Competition and X-Efficiency: A Reply," *Journal of Political Economy*, May/June 1973, 81, 765-77.
- \_\_\_\_\_, (1973b) "Notes on X-Efficiency and Technical Progress," in B. Eliezer Ayal, ed., *Micro Aspects of Development*, New York: Praeger, 1973, 18-40.
- \_\_\_\_\_, "Aspects of the X-Efficiency Theory of the Firm," *Bell Journal of Economics*, August 1975, 6, 580-606.
- \_\_\_\_\_, *Beyond Economic Man: A New Foundation for Micro-Economics*, Cambridge: Harvard University Press, 1976.
- \_\_\_\_\_, (1977a) "X-Efficiency, Technical Efficiency, and Incomplete Information Use: A Comment," *Economic Development and Cultural Change*, January 1977, 25, 311-16.
- \_\_\_\_\_, (1977b) "X-Efficiency Theory, Conventional Entrepreneurship, and Excess Capacity Creation in LDCs," *Economic Development and Cultural Change*, Supplement, 1977, 25, 288-99.
- \_\_\_\_\_, (1978a) *General X-Efficiency Theory & Economic Development*, New York and London: Oxford University Press, 1978.
- \_\_\_\_\_, (1978b) "X-Inefficiency Xists—Reply to an Xorcist," *American Economic Review*, March 1978, 68, 203-11.
- \_\_\_\_\_, (1978c) "On the Basic Proposition of X-Efficiency Theory," *American Economic Review Proceedings*, May 1978, 68, 328-34.
- \_\_\_\_\_, (1979a) "A Branch of Economics is Missing: Micro-Micro Theory," *Journal of Economic Literature*, June 1979, 17, 477-502.
- \_\_\_\_\_, (1979b) "X-Efficiency: From Concept to Theory," *Challenge*, September/October 1979, 22, 13-22.
- Libecap, Gary D., "Economic Variables and the Development of Law: The Case of Western Mineral Rights," *Journal of Economic History*, June 1978, 38, 338-62.
- McNulty, Paul J., "Allocative Efficiency vs. 'X-Efficiency': Comment," *American Economic Review*, December 1967, 57, 1249-52.
- Manne, Henry G., "Mergers and the Market for Corporate Control," *Journal of Political Economy*, April 1965, 73, 753-61.
- Marris, Robin, *The Economics of 'Managerial' Capitalism*, New York: Free Press of Glencoe, 1964.
- \_\_\_\_\_, and Mueller, Dennis C., "The Corporation, Competition and the Invisible Hand," *Journal of Economic Literature*, March 1980, 18, 32-63.
- Nagel, Ernest, "Assumptions in Economic Theory," *American Economic Review Pro-*

- ceedings, May 1963, 53, 211-19.
- Nicols, Alfred, "Stock versus Mutual Savings and Loan Associations: Some Evidence of Differences in Behavior," *American Economic Review Proceedings*, May 1967, 57, 337-47.
- Ohlin, Goran, "Review of *Productivity and Profitability: Studies of the Role of Capital in the Swedish Economy* by E. Lundberg," *American Economic Review*, September 1962, 52, 827-29.
- Oi, Walter Y., "Heterogeneous Firms and the Organization of Production," mimeo., March 1981.
- Olsen, E. Odgers, Jr., "The Effort Level, Work Time, and Profit Maximization," *Southern Economic Journal*, April 1976, 42, 644-52.
- Parish, Ross M., "The Efficiency Effects of Trade and Protection: Comment," in I. A. McDougall and R. H. Snape, eds., *Studies in International Economics*, Amsterdam-London: North-Holland, 1970, 11-14.
- \_\_\_\_\_ and Ng, Yew-Kwang, "Monopoly, X-Efficiency and the Measurement of Welfare Loss," *Economica*, August 1972, 39, 301-08.
- Penrose, Edith T., *The Theory of the Growth of the Firm*, New York: John Wiley and Sons, 1959.
- Posner, Richard A., "The Social Cost of Monopoly and Regulation," *Journal of Political Economy*, August 1975, 83, 807-27.
- Primeaux, Walter J., "An Assessment of X-Efficiency Gained through Competition," *Review of Economics and Statistics*, February 1977, 59, 105-8.
- Radice, H. K., "Control Type, Profitability and Growth in Large Firms," *Economic Journal*, September 1971, 81, 547-62.
- Reder, Melvin W., "Chicago Economics: Permanence and Change," *Journal of Economic Literature*, March 1982, 20, 1-38.
- Rostas, Laszlo, *Comparative Productivity in British and American Industry*, Cambridge: National Institute of Economic Sociology, 1964.
- Salter, W. E. G., *Productivity and Technical Change*, Cambridge: Cambridge University Press, 1960.
- Schlegel, R., *Completeness in Science*, New York: Appleton-Century-Crofts, 1967.
- Schwartzman, David, "The Burden of Monopoly," *Journal of Political Economy*, December 1960, 68, 727-29.
- \_\_\_\_\_, "Competition and Efficiency: Comment," *Journal of Political Economy*, May/June 1973, 81, 756-64.
- Scitovsky, Tibor, "A Note on Profit Maximization and Its Implications," *Review of Economic Studies*, No. 1, 1943, 11, 57-60.
- \_\_\_\_\_, *Economic Theory and Western Economic Integration*, Stanford: Stanford University Press, 1958.
- Shapiro, Kenneth H. and Müller, Jürgen, "Sources of Technical Efficiency: The Role of Modernization and Information," *Economic Development and Cultural Change*, January 1977, 25, 293-310.
- Shelton, John P., "Allocative Efficiency vs. 'X-Efficiency': Comment," *American Economic Review*, December 1967, 57, 1252-58.
- Shen, Tsung-Yuen, "Technology Diffusion, Substitution, and X-Efficiency," *Econometrica*, March 1973, 41, 263-84.
- Shepherd, William G., "The Elements of Market Structure," *Review of Economics and Statistics*, February 1972, 54, 25-37.
- Simon, Herbert A., "Theories of Decision Making in Economics and Behavioral Science," *American Economic Review*, June 1959, 49, 253-83.
- \_\_\_\_\_, "New Developments in the Theory of the Firm," *American Economic Review Proceedings*, May 1962, 52, 1-15.
- \_\_\_\_\_, "Rational Decision Making in Business Organizations," *American Economic Review*, September 1979, 69, 213-25.
- Stigler, George J., "The Economics of Information," *Journal of Political Economy*, June 1961, 69, 213-25.
- \_\_\_\_\_, "The Existence of X-Efficiency," *American Economic Review*, March 1976, 66, 213-16.
- Stiglitz, Joseph E., "Incentives, Risk, and Information: Notes Toward a Theory of Hierarchy," *Bell Journal of Economics*, Autumn 1975, 6, 552-79.
- Tullock, Gordon, "The Welfare Costs of Tariffs, Monopolies, and Theft," *Western Economic Journal*, June 1967, 5, 224-32.
- Umbeck, John, "Might Makes Rights: A Theory of the Foundation and Initial Distribution of Property Rights," *Economic In-*

- quiry, January 1981, 19, 38-59.
- Veblen, Thorstein, "Why Is Economics Not An Evolutionary Science?," in his *The Place of Science in Modern Civilization*, New York: B. W. Huebsch, 1919, 56-81.
- Williamson, Oliver E., "Managerial Discretion and Business Behavior," *American Economic Review*, December 1963, 53, 1032-57.
- , *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*, Englewood Cliffs: Prentice-Hall, 1964.
- , "Hierarchical Control and Optimum Firm Size," *Journal of Political Economy*, April 1967, 75, 123-38.
- , *Corporate Control and Business Behavior*, Englewood Cliffs: Prentice-Hall, 1970.
- , *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: Free Press, 1975.
- , "Transaction Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October 1979, 22, 233-61.
- , "The Modern Corporation: Origins, Evolution, Attributes," *Journal of Economic Literature*, December 1981, 19, 1537-68.
- International Labor Organization, *Payment by Results*, Geneva: ILO Studies and Reports, 1951.

# On the Misuse of Accounting Rates of Return to Infer Monopoly Profits

By FRANKLIN M. FISHER AND JOHN J. MCGOWAN\*

Accounting rates of return are frequently used as indices of monopoly power and market performance by economists and lawyers.<sup>1</sup> Such a procedure is valid only to the extent that profits are indeed monopoly profits, accounting profits are in fact economic profits, and the accounting rate of return equals the economic rate of return.

The large volume of research investigating the profits-concentration relationship uniformly relies on accounting rates of return, such as the ratio of reported profits to total assets or to stockholders' equity as the measure of profitability to be related to concentration.<sup>2</sup> Many users of accounting rates of return seem well aware that profits as reported by accountants may not be consistent from firm to firm or industry to industry and may not correspond to economists' definitions of profits. Likewise, they recognize that accountants' statements of assets, hence also stockholders' equity, may fail to correspond to economically acceptable definitions, because accounting practices do not provide for the capitalization of certain activities such as research and development and do not incorporate al-

lowances for inflation. This is to say they are well aware of certain measurement problems which arise in using available accounting information to measure profitability. They seem, however, totally unaware of a much deeper conceptual problem, namely, that accounting rates of return, even if properly and consistently measured, provide almost no information about economic rates of return.<sup>3</sup>

The economic rate of return on an investment is, of course, that discount rate that equates the present value of its expected net revenue stream to its initial outlay. Putting aside the measurement problems referred to above, it is clear that it is the economic rate of return that is equalized within an industry in long-run industry competitive equilibrium and (after adjustment for risk) equalized everywhere in a competitive economy in long-run equilibrium. It is an economic rate of return (after risk adjustment) above the cost of capital that promotes expansion under competition and is produced by output restriction under monopoly. Thus, the economic rate of return is the only correct measure of the profit rate for purposes of economic analysis.<sup>4</sup> Accounting rates of return are useful only insofar as they yield information as to economic rates of return.<sup>5</sup>

\*Fisher is professor of economics, Massachusetts Institute of Technology. McGowan was Vice-President, Charles River Associates. He died on April 7, 1982. This paper is based on work done for Fisher's testimony as a witness for IBM in *U.S. v. IBM* (69 Civ. 200, U.S. District Court, Southern District of New York). We are indebted to Larry Brownstein, Steven Hendrick, and especially Karen Larson and Leah Hutten for computational and programming assistance. Any errors are our responsibility.

<sup>1</sup>Aside from *U.S. v. IBM*, see, for example, Joseph Cooper, p. 15; the various industry studies in Walter Adams; and the discussion in Philip Areeda and Donald Turner, Vol. II, pp. 331-41.

<sup>2</sup>See the comprehensive reviews of this literature by Leonard Weiss and more recently by F. M. Scherer, pp. 267-95. Additional accounting problems raised by attempting to measure profitability by line of business are discussed extensively in George Benston.

<sup>3</sup>A referee suggests that even the crudest accounting information tells us IBM is more profitable than American Motors (AMC), but we disagree. Surely accounting information tells us IBM generates more dollars of profits per dollar of assets than does AMC but, as the examples below demonstrate, that information alone does not tell us which firm is more profitable in the sense of having a higher economic rate of return.

<sup>4</sup>This is literally true only if the cost of capital is first subtracted. In what follows below, we follow the usual empirical practice of measuring all rates of return before such subtraction.

<sup>5</sup>The existence of a uniquely defined economic rate of return—which we now assume for the theoretical analysis below and which occurs in all the examples—is

Now, it should be obvious that only by the merest happenstance will the accounting rate of return on a given investment, taken as the ratio of net revenue to book value in a particular year,<sup>6</sup> be equal to the economic rate of return that makes the present value of the entire net revenue stream equal to the initial capital cost. Indeed, as we shall see below, accounting rates of return on individual investments generally vary all over the lot. Hence, only if such fluctuations are somehow averaged out by a firm's investment behavior over time will its accounting rate of return even be roughly constant—let alone approximate the economic rate of return.<sup>7</sup>

It is easy to show that such averaging requires that the firm grow exponentially, investing in the same mix of investment types each year—an investment type being defined by a time shape of net revenues. Even in such an unrealistically favorable case, the accounting rate of return will generally depend on the rate of growth, equalling the economic rate of return only by accident. Furthermore, the relationship between the accounting and economic rates of return depends on the time shape of net revenues.

---

guaranteed only if the net revenue stream stemming from an investment has any negative terms occurring before the positive ones. If the economic rate of return fails to be unique, then, while present value calculations using the cost of capital remain the correct method for analyzing profitability, profitability cannot be summarized correctly by any rate of return, including accounting rates of return.

<sup>6</sup>Throughout this paper we work with accounting rates of return defined as ratios of profits to book values of capital. Similar (but not identical in detail) results apply to accounting rates of return on stockholders' equity. The precise relations involved can, in principle, be inferred from the results given below. (Such results do apply directly to accounting rates of return on stockholders' equity even in detail if we consider the firms being analyzed to hold neither debt nor retained earnings.)

<sup>7</sup>For discussion purposes—and in our examples below—we assume that the firm achieves the same economic rate of return on all its investments, and thus speak of “the” economic rate of return for the firm without worrying about differences between average and marginal rates. This is, of course, the most favorable case for the accounting rate of return for the firm as a whole.

Hence, only by accident will accounting rates of return be in one-to-one correspondence with economic rates of return. We show by example below that the effects involved cannot be assumed to be small—indeed, they can be large enough to account for the entire interfirm variation in accounting rates of return among the largest firms in the United States.

The plan of the paper is as follows. Section I summarizes the theoretical results which are proved and elucidated in the Appendix. These results establish the relationships among the various rates of return, time shapes, and rates of growth, and demonstrate in principle that accounting rates of return are not informative. The balance of the paper analyzes a series of relatively simple examples to show that the theoretical effects are not so small that they can be neglected in practice. Indeed, they are very large. A ranking of firms by accounting rates of return can easily invert a ranking by economic rates of return.

Before proceeding, we note that some of the theoretical results given below are not new. Ezra Solomon wrote a number of articles culminating in one dealing with the case of exponential growth in 1970. Thomas Stauffer published various theorems a year later (1971) and also attempted to make adjustments to accounting rates of return to correct for alternative cash flow profiles in testimony for the FTC in the *Ready to Eat Cereal Litigation*.<sup>8</sup> J. Leslie Livingstone and Gerald Salamon (1971) have also studied and attempted to determine a relationship between the accounting and internal rates of return. Yet, perhaps because Solomon's focus was on the correct concepts of rate of return and cost of capital for rate regulation, or perhaps because none of the studies cited makes clear just how large the effects involved can be, the importance of these matters for more general industrial organization research appears to have gone largely unnoticed. It is our hope that the self-contained discussion of the present paper and, especially, the mag-

<sup>8</sup>The proofs given below are different from Stauffer's proofs, and, we think, more suitable for our present purposes than his where the propositions coincide.

nitudes of the effects exhibited in the examples below will remedy this.

### I. Summary of Theoretical Results

The main theoretical results, which are proved and elucidated in the Appendix, are as follows:

(a) Unless depreciation schedules are chosen in a particular way, so that the value of the investment is calculated as the present value at the economic rate of return of the stream of benefits remaining in it<sup>9</sup>—a choice which is exceptionally unlikely to be made—the accounting rate of return on a particular investment will differ from year to year, and will not in general equal the economic rate of return on that investment in any year.

(b) The accounting rate of return for the firm as a whole will be an average of the accounting rates of return for individual investments made in the past. The weights in that average will consist of the book value of those different investments which in turn depend on the depreciation schedule adopted, and, particularly, on the amount and timing of such investments.

(c) Unless the proportion of investments with a given time shape remains fixed every year, and unless the firm simply grows exponentially, increasing investments in each and every type of asset<sup>10</sup> by the same proportion for every year, the accounting rate of return to the firm as a whole cannot even be expected to be constant, let alone be equal to the economic rate of return.

(d) Even where the firm does operate in such an unrealistic manner—the case most favorable to the accounting rate of

return—the accounting rate of return will vary with the rate of growth of the firm, and will not generally equal the economic rate of return.

(e) The only reliable inferences concerning the economic rate of return that can be drawn (and only in such an unrealistically favorable case) from examination of the accounting rate of return stem from the fact that the accounting rate of return and the economic rate of return will be on the same side of the firm's exponential growth rate. If the accounting rate of return is higher than the growth rate, then the economic rate of return is also higher than the growth rate. If the accounting rate of return is lower than the growth rate, then the economic rate of return is lower than the growth rate. If the accounting rate of return equals the growth rate, and in this case *alone*, the economic rate of return is guaranteed to be equal to the accounting rate of return.<sup>11</sup>

(f) Even in the unrealistically favorable exponential growth case, the accounting rate of return depends *crucially* on the time shape of benefits, and the effect of growth on the accounting rate of return also depends on that time shape. In particular, it is not true that rapidly growing firms tend to understate their profits and slowly growing firms tend to overstate them. The effect can go the other way.<sup>12</sup>

(g) All these results apply both to before- and after-tax rates of return.

### II. The Likely Size of the Effects

We now show by example that differences between the accounting and economic rates of return can be quite large indeed. For the sake of economy we examine only differences in after-tax rates of return. We as-

<sup>9</sup>Such a "natural" depreciation formula—which we shall term "economic depreciation"—was first suggested by Harold Hotelling in 1925. It is somewhat misleading, however, to say that the fundamental conceptual problems discussed in the present paper are basically matters of depreciation accounting. Rather, there exists a particular form of depreciation which will correct those problems which stem from a fundamental difference between the economic and accounting rates of return. These problems arise even where machines never wear out. An example is given in Fisher (1979).

<sup>10</sup>Two assets are said to be of the same "type" if they yield the same time shape of benefits.

<sup>11</sup>It is worth pointing out that these results apply to accounting rates of return on total assets, not directly to accounting rates of return on stockholders' equity. Further, they apply to accounting rates of return on beginning-of-year, not end-of-year or yearly average assets. As the examples below show, the problem of making inferences from accounting rates of return on end-of-year (or average) assets is even worse—if possible—than when beginning-of-year assets are used.

<sup>12</sup>Compare Cooper, pp. 132–33.

TABLE 1—AFTER-TAX ACCOUNTING RATES OF RETURN<sup>a</sup>  
(Percent for the Q-Profile; Six-Year Life; No Delay)

Year	Gross Profits (Cash Flow Before-Tax)	Depreciation	After-Tax Profits	Beginning- of-Year Assets		End-of-Year Assets	
				Net	Accounting Rate of Return	Net	Accounting Rate of Return
1	23.3	28.6	(5.3)	100.0	(5.3)	71.4	(7.4)
2	44.1	23.8	11.2	71.4	15.7	47.6	23.5
3	51.9	19.0	18.1	47.6	38.0	28.6	63.3
4	40.5	14.3	14.4	28.6	50.3	14.3	100.7
5	20.2	9.5	5.9	14.3	41.3	4.8	122.9
6	7.8	4.8	1.7	4.8	35.4	0	Infinite

<sup>a</sup>Tax rate: 45 percent; After-tax economic rate of return: 15 percent; Sum-of-the-years' digits depreciation.

sume a corporate tax rate of 45 percent, and (for most examples) fix the after-tax economic rate of return at 15 percent while varying growth rates and depreciation methods and the time shape of benefits.<sup>13</sup> Enormous variations in the accounting rates of return are readily generated.

#### A. The "Q-Profile"

We start with an investment whose benefits begin immediately and last for six years, and follow the time shape exhibited in column 2 of Table 1. For convenience we refer to this shape as the Q-profile.<sup>14</sup> The figures in column 2 are scaled to produce an after-tax economic rate of return of 15 percent on an

initial investment of \$100 when sum-of-the-years' digits depreciation over a six-year life is used. The remainder of the table shows the calculation of the corresponding accounting rate of return each year.

Plainly, the after-tax accounting rates of return vary substantially. They never equal the after-tax economic rate of return (15 percent), and exceed it in every year with positive net profits. Real-life firms do not generally exhibit such variation in their accounting rates of return because the averaging effects of growth, as it were, attribute profits from past investment to the book value of investments whose profit results are yet to come, rather than to the declining book value of such past investment.

While such an averaging effect tends to stabilize the accounting rate of return, it becomes a hodgepodge devoid of information about the economic rate of return. This point is illustrated by Table 2, which presents asymptotic accounting rates of return assuming constant exponential growth for three different versions of the Q-profile, each with the same tax rate (45 percent) and after-tax economic rate of return (15 percent).<sup>15</sup> The first version (the case of Table 1)

<sup>13</sup>Fifteen percent was roughly the average accounting rate of return in U.S. manufacturing corporations in 1978 (*Economic Report of the President*, 1979, pp. 279-91). If accounting and economic rates of return tended to coincide, 15 percent would be a reasonable choice for the economic rate of return. Since the rates do not generally coincide, the choice is immaterial. Choosing a lower economic rate of return would reduce the range of accounting rates of return in the results below (for the same examples), but would not affect the conclusions.

With a fixed capital investment, a given time shape of gross profits before depreciation and taxes results in different after-tax economic rates of return for different depreciation methods. To fix the after-tax economic rate of return for a given time shape, therefore, we adjust the height of the gross profit benefit stream proportionally to produce the desired after-tax economic rate of return.

<sup>14</sup>This shape was (erroneously) suggested during U.S. v. IBM as being typical of IBM's experience. We use it for convenience.

<sup>15</sup>In this context, exponential growth takes place by repeated investment in the same type of project; i.e., all investments have the same time-shape of benefits. This is obviously an unrealistic assumption, but one which is more likely to produce equality between accounting and economic rates of return than more realistic assumptions.

TABLE 2—ASYMPTOTIC ACCOUNTING RATES OF RETURN (%) ON THREE VERSIONS OF THE Q-PROFILE<sup>a</sup>

Growth Rate	Six-Year Life (No Delay)			Seven-Year Life (One-Year Delay)			Eight-Year Life (Two-Year Delay)		
	Straight Line	Declining Balance	Sum-of-Years' Digits	Straight Line	Declining Balance	Sum-of-Years' Digits	Straight Line	Declining Balance	Sum-of-Years' Digits
A. Beginning-of-Year Assets									
0	15.2	17.8	18.1	18.1	21.3	22.0	21.0	24.7	25.9
5	15.2	16.9	17.0	17.0	19.1	19.4	18.9	21.1	21.7
10	15.1	15.9	15.9	16.0	17.0	17.1	16.9	17.9	18.1
15	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0
20	14.8	14.1	14.1	14.0	13.2	13.1	13.3	12.4	12.3
25	14.7	13.3	13.3	13.1	11.5	11.4	11.7	10.1	9.9
30	14.5	12.5	12.6	12.2	10.0	9.9	10.3	8.0	7.8
B. End-of-Year Assets									
0	15.2	17.8	18.1	18.1	21.3	22.0	21.0	24.7	25.9
5	14.5	16.1	16.2	16.2	18.1	18.5	18.0	20.1	20.7
10	13.7	14.5	14.5	14.6	15.4	15.5	15.3	16.3	16.5
15	13.0	13.0	13.0	13.0	13.0	13.0	13.0	13.0	13.0
20	12.4	11.8	11.8	11.7	11.0	10.9	11.1	10.3	10.2
25	11.7	10.6	10.7	10.5	9.2	9.2	9.4	8.1	7.9
30	11.1	9.6	9.7	9.4	7.7	7.6	7.9	6.2	6.0

<sup>a</sup>See Table 1.

has no delay between investment and the beginning of the benefit stream, and depreciation is taken over the resulting six-year life. The second version has a seven-year life including a one-year's delay between investment and initial return. The third has an eight-year life including a two-year delay between investment and initial return. Except for the lag at the beginning and differences in scale, the gross benefit stream is the same in each case. Panel A of the table gives accounting rates of return on beginning-of-year assets; Panel B gives those on end-of-year assets.

Several things are apparent from Table 2. First, the accounting rates of return only equal the economic rate of return of 15 percent when the growth rate is also 15 percent and when the accounting rate of return is measured on beginning-of-year assets. Otherwise, the accounting rates vary from seven points below to almost eleven points above the economic rate of return.

Second, it is not true (as is sometimes stated) that more rapid depreciation, other things equal, tends to understate accounting rates of return. In this example, when the rate of growth is below 15 percent, declining balance and sum-of-the-years' digits depreciation produces a higher accounting rate of return than straightline depreciation for given

growth rates, time profiles, and economic rates of return. The effect is reversed when the growth rate exceeds the economic rate of return of 15 percent. This illustrates a general proposition: more rapid depreciation *increases* the accounting rate of return (measured on beginning-of-year assets) when the growth is less than the economic rate of return, and *decreases* the accounting rate of return when the growth rate exceeds the economic rate of return.<sup>16</sup> Since this is the only point about depreciation which we wish to demonstrate, we provide only results for sum-of-the-years' digits depreciation in the rest of this paper.<sup>17</sup>

In all the examples in Table 2, firms growing at rates greater than the economic rate of

<sup>16</sup>By Theorem 1, the changeover point is also where the growth rate equals the accounting rate of return on beginning-of-year assets.

<sup>17</sup>There is one additional point about depreciation which we shall not bother to exemplify. Since the depreciation method chosen affects the time shape of the after-tax benefit stream, the relationship of after-tax accounting rates to the growth rate is particularly sensitive to the depreciation method. It can even happen that faster growth increases accounting rates of return for one choice of depreciation method and decreases them for another—all for the same pre-tax benefit time shape and the same after-tax economic rate of return. This makes adjustments for growth even harder to make than appears from the examples below.

TABLE 3—ASYMPTOTIC ACCOUNTING RATES OF RETURN (%) ON FOUR VERSIONS OF THE Q-PROFILE<sup>a</sup>

Growth Rate	Ten-Year Life (No Delay, Last Year Spread)	Six-Year Life (No Delay)	Seven-Year Life (One-Year Delay)	Eight-Year Life (Two-Year Delay)
A. Beginning-of-Year Assets				
0	13.9	18.1	22.0	25.9
5	14.5	17.0	19.4	21.7
10	14.8	15.9	17.1	18.1
15	15.0	15.0	15.0	15.0
20	15.1	14.1	13.1	12.3
25	15.1	13.3	11.4	9.9
30	15.0	12.6	9.9	7.8
B. End-of-Year Assets				
0	13.9	18.1	22.0	25.9
5	13.8	16.2	18.5	20.7
10	13.5	14.5	15.5	16.5
15	13.0	13.0	13.0	13.0
20	12.6	11.8	10.9	10.2
25	12.0	10.7	9.2	7.9
30	11.5	9.7	7.6	6.0

<sup>a</sup>See Table 1.

return of 15 percent have accounting rates of return on beginning-of-year assets less than the economic rate of return, while those growing at rates less than the economic rate of return all have accounting rates of return on beginning-of-year assets greater than the economic rate of return.<sup>18</sup> Contrary to what might be expected, this qualitative relationship provides no practical basis for adjusting accounting rates of return so that they will accurately reflect economic rates of return.

Table 2, for example, shows that firms which use sum-of-the-years' digits depreciation and grow at 5 percent have accounting rates of return on beginning-of-year assets which range from 17.0 to 21.7 percent. Thus, even for firms with the same growth rate and depreciation method, the required adjustment varies from 2 to 6.7 percentage points depending upon the time profile. Clearly, the time profile, depreciation method, and growth rate must all be known before accounting rates of return can be adjusted to reflect economic rates of return.

In the foregoing examples, for a given time shape, faster-growing firms have lower

accounting rates of return than slower-growing ones with the same economic rate of return. We have seen that even if this were a universal phenomenon, it would not provide a way to adjust accounting rates of return to reflect economic rates of return, since different firms will generally have different time shapes and therefore require different adjustments. The difficulties are even worse in practice, because the accounting rate of return can actually *rise* with the growth rate, causing *slower*-growing firms to have their economic rates of return *understated*. Thus, even the strong assumption that firms have the same time profile is insufficient to permit adjustment of accounting rates of return; the specific profile must also be known in order to make inferences about the ranking of economic rates of return.

We demonstrate this phenomenon by taking the original Q-profile (six-year life and no delay) and spreading the last year's gross profits out evenly over five years (years 6–10) instead of having them all in year 6. Table 3 shows that this small change in the profile produces an increasing relationship between the growth rate and the accounting rate of return. The original results for sum-of-the-year's digits depreciation are reproduced for ease of comparison.

<sup>18</sup>So simple a relationship does not hold if the accounting rate of return is based on end-of-year assets.

Focusing on the first column (10-Year Life), we see that the accounting rate of return on beginning-of-year assets actually begins by rising with the growth rate, reaching the value of the economic rate of return (as it must) at a 15 percent growth rate, and then going slightly above it before falling back again. (It is a special feature of this particular example that these values are all close to the economic rate of return of 15 percent.) The behavior of the accounting rate of return on end-of-year assets is different. This magnitude falls with the growth rate (in this example), but it exhibits still another phenomenon. As opposed to the previous example, where the accounting rates of return on both beginning- and end-of-year assets were above the economic rate of return of 15 percent for low growth rates and below it for large ones, here the accounting rate of return on end-of-year assets starts *and finishes* below the economic rate of return of 15 percent. There is *no* rate of growth for which the accounting rate of return on end-of-year assets equals the economic rate of 15 percent.

The impossibility of making inferences about relative profit rates should be obvious even within the confines of these examples, all of which represent only relatively slight variations on the same profile. *Every one of the firms exhibited in Table 3 has the same underlying after-tax economic rate of return. Yet their after-tax accounting rates of return on end-of-year assets vary from 6.0 to 25.9 percent.*<sup>19</sup>

Further, it is impossible to infer anything about relative profitability by attempting to adjust for growth rates. For example, each row of Table 3 involves firms with the same growth rate, so that there is nothing to adjust for in comparing them; yet, except for the special row corresponding to the point where the growth rate is equal to the true after-tax economic rate of return, the after-tax accounting rates of return continue to vary. For the row corresponding to 5 percent growth, for example, after-tax accounting

TABLE 4—BEFORE-TAX BENEFIT STREAMS FROM AN INVESTMENT OF \$100<sup>a</sup>

Year	X Firm (\$)	Y Firm (\$)
1	90.2	107.0
2	27.1	10.7
3	18.0	10.7
4	9.0	10.7
5	9.0	10.7
6	9.0	10.7

<sup>a</sup>See Table 1.

rates of return vary between 13.8 and 20.7 percent. For the row corresponding to 25 percent, they vary between 7.9 and 12.0 percent. Further, it is not correct to say that slow-growing firms have accounting rates of return that overstate their economic rate, while fast-growing firms have accounting rates of return that understate them. Continuing to use accounting rates of return on end-of-year assets, the firm just introduced (10-Year Life) has an accounting rate of return which understates its economic rate of return at all levels of growth. If one uses beginning-of-year assets, it has accounting rates of return which tend to understate its economic rate of return at low rates of growth and (slightly) overstate it at higher ones.

Moreover, the phenomenon of accounting rates of return increasing with the growth rate can be considerably more marked if we use other profiles. Table 4 shows the before-tax benefit stream (corresponding to an initial investment of \$100, an economic rate of return of 15 percent, and sum-of-the-years' digits depreciation over a six-year life) for two other profiles (*X* firm and *Y* firm). Table 5 shows the after-tax accounting rates of return for these firms when they grow exponentially at various rates. The after-tax accounting rates of return on beginning-of-year assets rise rather rapidly with the growth rate. The after-tax accounting rate of return on end-of-year assets also rises with the growth rate. However, as was also the case for the variation on the *Q*-profile examined earlier, it does not rise by enough to get to the economic rate of return of 15 percent.

<sup>19</sup>Here and later, the results for beginning-of-year assets are similar.

TABLE 5—ASYMPTOTIC ACCOUNTING RATES OF RETURN (%)  
FOR X-FIRMS AND Y-FIRMS<sup>a</sup>

Growth Rate	Beginning-of-Year Assets		End-of-Year Assets	
	X Firm	Y Firm	X Firm	Y Firm
0	12.9	12.5	12.9	12.5
5	13.6	13.3	13.0	12.7
10	14.3	14.2	13.0	12.9
15	15.0	15.0	13.0	13.0
20	15.7	15.8	13.0	13.2
25	16.3	16.6	13.0	13.3
30	16.9	17.3	13.0	13.3

<sup>a</sup>See Table 1.

### III. Conclusions

That the accounting rate of return—after tax as well as before tax—is a misleading measure of the economic rate of return is evident from examining cases of single projects such as in Table 1. The cases shown in later tables are unduly *favorable* to the accounting rate of return in that they mask its behavior by averaging. That averaging effect is achieved by the quite unrealistic assumption that investment by the firm always brings in the same time shape of returns, and that the firm grows each year by increasing its investments at the same percentage rate. Even on such favorable terms, it is impossible to infer either the magnitude or direction of differences in economic rates of return from differences in accounting rates of return. This is because such inferences require not only correction for growth rates, but *also* knowledge of the time shapes of returns.

The level and behavior of the accounting rate of return are both sensitive to the type of time shape used. Even within the *Q*-profile example, the rates vary depending on when the time shape begins and how the last few years are spread out. There is every reason to suppose that firms differ in the time shapes of their investments, and that a particular firm's investments will also differ among themselves. Thus, comparisons of accounting rates of return to make infer-

ences about monopoly profits is a baseless procedure.

This conclusion can be most dramatically demonstrated by juxtaposing accounting rates of return for firms with different time shapes and *different* economic rates of return. When this is done, it is easy to see that firms with *higher* accounting rates of return can have *lower* economic rates of return. Table 6 gives after-tax economic rates of return and after-tax accounting rates of return on end-of-year assets for three growth rates (0, 5, and 10 percent), and for each of the six time shapes already discussed as well as two other "one-hoss shay" time shapes.<sup>20</sup> For *each* growth rate, the examples are chosen so that the eight firms represented are ranked in *ascending* order of economic rates of return and in *descending* order of accounting rates of return—a complete reversal even with growth rates constant.

Examination of Table 6 shows again that no inference about relative after-tax economic rates of return is possible from after-tax accounting rates of return. For example, the lowest after-tax economic rate of return in the table is that for the *Q*-profile with an eight-year life at a zero growth rate. For that firm, the after-tax economic rate of return is 13 percent. Yet, its after-tax accounting rate

<sup>20</sup>The one-hoss shay time shapes have a constant return (no lag) for four and six years, respectively, and zero returns thereafter.

TABLE 6—AFTER-TAX ECONOMIC RATES OF RETURN (*E*) AND ASYMPTOTIC ACCOUNTING RATES OF RETURN ON END-OF-YEAR ASSETS (*A*) FOR EIGHT TIME SHAPES<sup>a</sup>

Growth Rate	Growth Rate					
	0 Percent		5 Percent		10 Percent	
	<i>E</i>	<i>A</i>	<i>E</i>	<i>A</i>	<i>E</i>	<i>A</i>
<i>Q</i> -Profile						
8-Year Life	13.0	21.6	16.0	22.6	17.8	21.2
(2-year delay)						
7-Year Life	14.0	20.2	17.0	21.6	18.8	20.9
(1-year delay)						
One-Hoss Shay						
6-Year Life	15.0	20.0	18.1	21.4	19.7	20.7
(no delay)						
4-Year Life	16.0	19.8	19.0	21.3	20.0	20.289
(no delay)						
<i>Q</i> -Profile						
6-Year Life	16.1	19.6	19.05	21.2	20.05	20.287
(no delay)						
10-Year Life	18.0	16.9	20.0	18.5	22.0	19.8
(no delay; last year spread)						
X Firm	19.0	16.2	21.0	17.8	23.0	19.2
Y Firm	19.2	15.8	21.2	17.4	23.2	18.9

<sup>a</sup>Tax rate: 45 percent; Sum-of-the-years' digits depreciation.

of return on end-of-year assets is 21.6 percent, the second *highest* accounting rate of return in the table, and a value well above that of 15.8 percent for the *Y* firm at zero growth, corresponding to a 19.2 percent economic rate of return. The 21.6 percent accounting rate of return so encountered is even above the 18.9 percent figure obtained for the *Y* firm at 10 percent growth—a figure which corresponds to an economic rate of return of 23.2 percent, the highest in the table, and more than 10 percentage points above the economic rate of return of 13 percent for the *Q*-profile with an eight-year life at zero growth. Similar examples of reversals occur throughout the table.

Nor can one eliminate these effects by correcting somehow for differences in rates of growth. The table as constructed exhibits a reversal of the ordering of economic and accounting rates of return with the rate of growth held constant. Rate of growth effects have thus *already* been removed from each pair of columns to an extent beyond that which one could hope to achieve in practice.

Moreover, it is not true that faster-growing firms should have their accounting rates of return adjusted upwards relative to slower growing ones. Consider the comparison between the *Q*-profile with a ten-year life at zero growth and the *Q*-profile with an eight-year life at 5 percent growth. The faster-growing firm has an accounting rate of return (22.6 percent) already greater than that of the slower-growing firm (16.9 percent), but its economic rate of return (16.0 percent) is *below* that of the slower-growing firm (18.0 percent). Adjusting the faster-growing firm's accounting rate of return *upwards* relative to that of the slower-growing one will make things *worse*, not better.

As all of this makes clear, there is no way in which one can look at accounting rates of return and infer anything about relative economic profitability or, a fortiori, about the presence or absence of monopoly profits. The economic rate of return is difficult—perhaps impossible—to compute for entire firms. Doing so requires information about both the past and the future which

outside observers do not have, if it exists at all.<sup>21</sup> Yet it is the economic rate of return which is the magnitude of interest for economic propositions. Economists (and others) who believe that analysis of accounting rates of return will tell them much (if they can only overcome the various definitional problems which separate economists and accountants) are deluding themselves. The literature which supposedly relates concentration and economic profit rates does no such thing, and examination of absolute or relative accounting rates of return to draw conclusions about monopoly profits is a totally misleading enterprise.

#### APPENDIX

##### I: BEFORE-TAX ANALYSIS

##### A. *The Accounting Rate of Return on Individual Investments*

We begin our analysis of the problem by considering the before-tax accounting and economic rates of return on a single investment. Later we shall consider the firm as being made up of a series of such investments which may be (but need not always be) of the same type. The after-tax case is treated below and shown to be isomorphic, although more complex.

An investment may be thought of for heuristic purposes as a "machine" costing one dollar. If this is invested at time 0, the firm experiences a stream of net benefits as a result. Such benefits include all changes in revenues and costs (other than the initial capital cost) which accrue to the firm as a result of making the investment. The flow of such benefits at time  $\theta$  is denoted by  $f(\theta)$ .<sup>22</sup>

<sup>21</sup>If one made the strong assumption that the same time shape of returns held for all investments made by a given firm throughout its life, then it might be possible to recover that time shape by regression of gross returns on a distributed lag of past investment. We are indebted to Zvi Griliches for this suggestion.

<sup>22</sup>The time origin is arbitrary. The flow of benefits is assumed to depend on the age of the machine only. Thus an investment at time  $t$  brings in benefits of  $f(\theta - t)$  at time  $\theta \geq t$ . Time dependence of the benefit stream can be handled below by thinking of it as equivalent to investment in different kinds of machines at different times.

The economic rate of return on a machine,  $r$ , is that discount rate which makes the discounted value of the benefit stream equal to the capital costs of the investment. In other words,  $r$  satisfies

$$(A1) \quad \int_0^{\infty} f(\theta) \exp(-r\theta) d\theta = 1.$$

We assume that the integral in (A1) is monotonically decreasing in  $r$  so that (A1) has a unique positive solution. This will be true if the negative portion of the net benefit stream (if any) precedes the positive portion. This is the usual case.<sup>23</sup>

Now the firm adopts a depreciation schedule for this machine. Let  $V(\theta)$  denote the book value of the machine as of time  $\theta$ . Then  $-V'(\theta)$  is the rate of depreciation at  $\theta$ , where the prime denotes differentiation. Plainly,  $V(0) = 1$ , and it makes sense to suppose that  $V(\infty) = 0$ , although this latter condition is not really needed.

Accounting profits attributable to this machine at time  $\theta$  will be equal to net benefits less depreciation. We can think of the accounting rate of return for this machine as the accounting rate of return which the firm would have if this were its only asset. Denoting that rate by  $b(\theta)$ ,

$$(A2) \quad b(\theta) = (f(\theta) + V'(\theta))/V(\theta).$$

The first question which comes immediately to mind is that of when  $b(\theta) = r$  for all  $\theta$  within the life of the machine. We prove this will occur if and only if the depreciation schedule adopted by the firm always values the machine as the discounted value of the future benefit stream, discounting at the economic rate of return,  $r$  (see Hotelling).

**THEOREM 1:**  $b(\theta) = r$  if and only if

$$(A3) \quad V(\theta) = \int_{\theta}^{\infty} f(u) \exp(-r(u - \theta)) du.$$

<sup>23</sup>If (A1) has more than one solution, then the economic rate of return is ill-defined and there is even less point in considering whether the accounting rate of return yields information about it.

PROOF:

(a) Suppose (A3) holds. Differentiating with respect to  $\theta$ , we obtain

$$(A4) \quad V'(\theta) = -f(\theta) + rV(\theta),$$

which when substituted in equation (A2) yields  $b(\theta) = r$ .

(b) Suppose  $b(\theta) \equiv r$ . Then, from (A2),

$$(A5) \quad V'(\theta) \equiv rV(\theta) - f(\theta).$$

This is a linear differential equation with an additive forcing function  $(-f(\theta))$ . Its solution is therefore in the form

$$(A6) \quad V(\theta) = C \exp(r\theta) + z(\theta),$$

where  $z(\theta)$  is any particular solution of (A5) and  $C$  is a constant to be determined by the initial conditions. However, by part (a) of the proof, the integral on the right-hand side of (A3) is a particular solution of (A5). Hence  $z(\theta)$  can be taken as that integral. Do this and note that  $z(0) = 1$  by (A1), the definition of the economic rate of return. Since we have  $V(0) = 1$ , setting  $\theta = 0$  in (A6) yields  $C = 0$ , and the theorem is proved.

Thus even where the firm has a single simple investment with no ambiguity about marginal vs. economic rates of return, the accounting rate of return will not equal the economic rate of return except for a particular choice of a depreciation schedule—which choice we may term “economic depreciation.”

The reason for this is not hard to find. The book value of the firm's assets reflects the investment expenditures made in the past less the depreciation already taken on them. The benefits for which such investments were made are at least partly in the future. Yet the accounting rate of return takes gross profits before depreciation as the benefit flow which happens to be currently occurring. Unless depreciation is chosen so as to reflect the change in *future* benefits in the appropriate way, there is no reason to suppose that such a calculation should equal the economic rate of return, and Theorem 1 shows that the two will generally not be equal.

Will firms tend to adopt an “economic depreciation” schedule yielding book value as in equation (A3)? This is pathologically unlikely. Except in the simple “Santa Claus” case of  $f(\theta) = k \exp(-\lambda\theta)$  which corresponds to exponential depreciation or other similarly special cases corresponding to straightline or other standard depreciation methods, the benefit stream from investment when plugged into (A3) will not yield depreciation schedules anything like those used by real-life firms to optimize after-tax profits given IRS rules or those schedules used for nontax purposes. Real investments will almost invariably have complicated time shapes for their benefit streams. Further, even relatively simple shapes yield economic depreciation schedules which are quite far from actual ones. To see this, one need only observe that if  $V(\theta)$  satisfies equation (A3), there is no reason that  $V'(\theta)$  must always be negative. Indeed, if the time stream of benefits starts low and then has a hump a few years out, taking economic depreciation would require writing up the value of assets for the first few years. Yet there is nothing bizarre about such an example.

We must, therefore, with pathologically unlikely exceptions, expect that the accounting rate of return on a particular machine,  $a(\theta)$ , will generally not equal the economic rate of return  $r$ . (How far off it can be is demonstrated by examples.) This should make us suspect that the same thing will generally be true of the firm as a whole, and we now go on to explore that question.

#### B. *The Accounting Rate of Return for the Firm as an Average*

It is fairly plain that the best hope for an accounting rate of return equal to the economic rate of return will occur if all investments made by the firm are exactly alike, since otherwise (as shown below), changes in the mix of investment types will change the accounting rate. So we begin by considering the case in which all machines are like the machine above.

It now becomes necessary to distinguish calendar time, denoted by  $t$ , from the age of a machine, denoted by  $\theta$ . We let  $I(t)$  be the

value of investment made at  $t$  (equals the number of machines purchased). Let  $K(t)$  denote the book value of the firm's assets at  $t$  and  $\pi(t)$  the value of its accounting profits at  $t$ . Then,

$$(A7) \quad K(t) = \int_{-\infty}^t I(u)V(t-u) du \\ = \int_0^{\infty} I(t-\theta)V(\theta) d\theta,$$

where  $\theta = t - u$ . Similarly,

$$(A8) \quad \pi(t) = \int_{-\infty}^t I(u)\{f(t-u) + V'(t-u)\} du \\ = \int_0^{\infty} I(t-\theta)\{f(\theta) + V'(\theta)\} d\theta \\ = \int_0^{\infty} I(t-\theta)V(\theta)b(\theta) d\theta,$$

using (A2).

Hence, letting  $a(t)$  be the firm's accounting rate of return at  $t$ :

$$(A9) \quad a(t) \equiv \pi(t)/K(t) \\ = [\int_0^{\infty} \{I(t-\theta)V(\theta)\}b(\theta) d\theta] \\ / [\int_0^{\infty} \{I(t-\theta)V(\theta)\} d\theta];$$

so that we have proved

**LEMMA 1:** *At any time  $t$ , the accounting rate of return for the firm as a whole is a weighted average of the individual accounting rates for its individual past investments, the weights being the book values of those past investments.*

It should be obvious that this result would also be true if machines were not always of one type.

We now ask whether such an average will equal the economic rate of return. First consider whether the average can even be independent of  $t$ . This can happen in two ways. First,  $b(\theta)$  might be independent of  $\theta$ . We know from Theorem 1 that this will happen for  $b(\theta) \equiv r$  only for the cases of economic depreciation already discussed which we rule

out. It is easy to show that  $b(\theta) \equiv q \neq r$  is impossible.<sup>24</sup>

The other way in which  $a(t)$  might be independent of  $t$  would be if the relative weights in the average did not change over time.<sup>25</sup>

$$(A10) \quad \frac{I(t_1 - \theta)V(\theta)}{I(t_2 - \theta)V(\theta)} = k,$$

whence

$$(A11) \quad \frac{I'(t_1 - \theta)}{I(t_1 - \theta)} = \frac{I'(t_2 - \theta)}{I(t_2 - \theta)},$$

for all  $(t_1, t_2)$ . Evidently it must then be the case that

$$(A12) \quad I(t) = M \exp(gt)$$

for some constant growth rate  $g$ .

The remainder of our investigation will concern the case of exponential growth with the scale factor,  $M$ , set equal to unity. This case is the most favorable to accounting rates of return approximating economic rates of return, since in its absence accounting rates of return will not even be constant, even though the economic rate of return is well defined and constant.

### C. The Effect of the Growth Rate in Exponential Growth

We are now dealing with a case in which the accounting rate of return is (at least

<sup>24</sup>To see this, observe that a proof essentially the same as that of Theorem 1 would show that  $b(\theta) \equiv q$  if and only if

$$(a) \quad V(\theta) = C \exp(q\theta) + z(\theta),$$

where

$$(b) \quad z(\theta) = \int_{\theta}^{\infty} f(u) \exp(-q(u-\theta)) du,$$

but  $V(0) = 1$  so that  $C = 1 - z(0) \neq 0$  if  $q \neq r$ . Then (A4) yields  $V(\infty) = \pm \infty$  depending on  $q \gtrless r$  and this is not possible.

<sup>25</sup>For a given distribution of  $b(\theta)$  there might be other possibilities, but these would be even more special than the case of economic depreciation already discussed. The statement in the text is true if  $a(t)$  is to be constant despite unknown variations in  $b(\theta)$  with  $\theta$ .

asymptotically) constant and given as

$$(A13) \quad a = \frac{\int_0^\infty \{ \exp(g(t-\theta)) V(\theta) \} b(\theta) d\theta}{\int_0^\infty \{ \exp(g(t-\theta)) V(\theta) \} d\theta}$$

where  $a$  denotes the (asymptotic) constant value. This is still a weighted average of the accounting rates of return on individual investments. Plainly, the growth rate  $g$  affects the weights. Since the accounting rates of return on individual investments will almost always not be constant in view of Theorem 1, changes in the weights will usually affect the average.

The present section studies such effects and asks, in particular, what inferences can be drawn concerning the economic rate of return  $r$ , from knowledge of the accounting rate of return  $a$ , and the growth rate  $g$ , without information on the time shape of benefits,  $f(\cdot)$ , since the latter information is plainly never available from the books of the firm—even assuming it is known in detail to the firm's forecasters.

The first thing to say in this regard is that while (as we shall show) there exist values of  $g$  for which  $a=r$ , these values will be the exception. One cannot expect accounting and economic rates of return to coincide even in the most favorable case of exponential growth and a single investment type except by the merest accident. What information *can* be gleaned from the accounting rate of return is analyzed in this section.

It will be convenient to set up the problem a little differently from the analyses above. Let  $\pi^*(t)$  denote the gross profits of the firm before depreciation. Let  $\delta(t)$  denote total depreciation taken at time  $t$ . Let  $K^*(t)$  denote the *undepreciated* value of the firm's capital stock. Let  $D(t)$  denote the total depreciation already taken on that stock. Finally, let  $a^* = \pi^*(t)/K^*(t)$ , so that  $a^*$  is the accounting rate of return which would be observed if there were no depreciation. The following relationships hold:

$$(A14) \quad a = \frac{\pi^*(t) - \delta(t)}{K^*(t) - D(t)},$$

$$\begin{aligned} (A15) \quad \pi^*(t) &\equiv \int_{-\infty}^t \exp(gu) f(t-u) du \\ &= \int_0^\infty \exp(g(t-\theta)) f(\theta) d\theta \\ &= \exp(gt) \int_0^\infty \exp(-g\theta) f(\theta) d\theta \\ &= \exp(gt) \pi^*(0), \end{aligned}$$

$$\begin{aligned} (A16) \quad K^*(t) &\equiv \int_{-\infty}^t \exp(gu) du \\ &= \exp(gt)/g \end{aligned}$$

$$\begin{aligned} (A17) \quad \delta(t) &= \int_{-\infty}^t \exp(gu) V'(t-u) du \\ &= \int_0^\infty \exp(g(t-\theta)) V'(\theta) d\theta \\ &= \exp(gt) \delta(0), \end{aligned}$$

$$\begin{aligned} (A18) \quad D(t) &= \int_{-\infty}^t \delta(u) du \\ &= \int_{-\infty}^t \exp(gu) \delta(0) du \\ &= \delta(0) \exp(gt)/g. \end{aligned}$$

Evidently, we have proved:

LEMMA 2:  $\delta(t)/D(t) = g$ .

We now study the effects of  $g$  on  $a^*$ .

LEMMA 3: (a) If  $g=r$ , then  $a^*=r=g$ .

(b)  $d \log a^*/d \log g < 1$ .

(c)  $a^*$  and  $r$  are always on the same side of  $g$ . That is,  $a^* < g \Leftrightarrow r < g$ ;  $a^* = g \Leftrightarrow r = g$ ;  $a^* > g \Leftrightarrow r > g$ .

PROOF:

(a) Using equations (A15) and (A16),

$$\begin{aligned} (A19) \quad a^* &= g \pi^*(0) \\ &\equiv g \int_0^\infty \exp(-g\theta) f(\theta) d\theta. \end{aligned}$$

If  $g = r$ , then  $\pi^*(0) = 1$  by the definition of the economic rate of return (A1), whence  $a^* = g = r$ .

(b) From (A19),

$$(A20) \quad \log a^* = \log g + \log \pi^*(0),$$

but examination of  $\pi^*(0)$  shows that it is necessarily decreasing in  $g$  since it is the discounted integral of future benefits from a single machine discounted at the rate  $g$ . Thus  $d \log a^* / d \log g < 1$ .

(c) These statements follow directly from (a) and (b).

Using Lemmas 2 and 3, we can now proceed to the main result of this section for the magnitude of interest, the accounting rate of return  $a$ , itself.

**THEOREM 2:**  *$a$  and  $r$  are always on the same side of  $g$ . That is,*

$$a < g \leftrightarrow r < g; \quad a = g \leftrightarrow r = g;$$

$$a > g \leftrightarrow r > g.$$

**PROOF:**

By definition,  $\pi^*(t) = a^* K^*(t)$ . By Lemma 2,  $\delta(t) = gD(t)$ . Substituting in (A14)

$$(A21) \quad a = \frac{a^* K^*(t) - gD(t)}{K^*(t) - D(t)} \gtrless g,$$

accordingly as  $a^* \gtrless g$ . The desired result now follows from Lemma 3.

A diagram may be illuminating here. In Figure 1, the growth rate is measured on the horizontal axis and rates of return on the vertical axis. The 45° line indicates where growth rates and rates of return are equal. Theorem 2 states that the accounting rate of return must be above the 45° line to the left of the dashed line at  $g = r$ ; it must pass through  $H$ , the point of intersection of the dashed line and the 45° line; and it must be below the 45° line to the right of the dashed line.

Can we say more than this? The answer is in the negative without information on the time shape of benefits  $f(\cdot)$ . In particular, it is

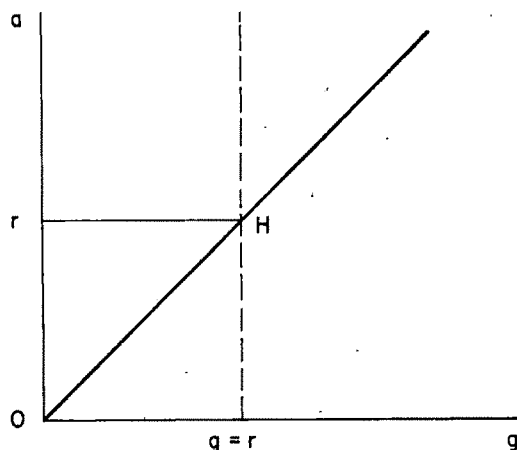


FIGURE 1

not the case that the direction of change of  $a$  with respect to  $g$  is signed. Nor is it true that  $r$  must lie between  $a$  and  $g$ . These facts are exemplified in the text.

## II. AFTER-TAX ANALYSIS

These same results apply to the analysis of the relationship between the after-tax economic rate of return and the after-tax accounting rate of return. This is obvious if the depreciation schedule used is not that used for tax purposes; in that case, the effect of taxes is just to change the benefit profile with the analysis the same as before, given the new benefit profile  $f(\cdot)$ . Moreover, the same thing is true if tax depreciation is used. To see this, let  $\alpha$  be the tax rate  $0 < \alpha < 1$  (assumed constant for simplicity). Let  $r'$  denote the after-tax economic rate of return. Then  $r'$  satisfies

$$(A22) \quad \int_0^\infty \{ (1 - \alpha)f(\theta) + \alpha d(\theta) \} \times \exp(-r'\theta) d\theta = 1,$$

where  $d(\theta)$  denotes depreciation on an asset of age  $\theta$  and  $f(\theta)$  denotes its *before-tax* benefits, as before.

This reflects the fact that the choice of a depreciation schedule,  $d(\cdot)$ , affects after-tax

returns. Define

$$(A23) \quad f^*(\theta) \equiv ((1-\alpha)f(\theta) + \alpha d(\theta)).$$

We now show that the analysis of the before-tax case applies directly to the after-tax case with  $f^*(\cdot)$ , the after-tax benefit schedule, replacing  $f(\cdot)$ , the before-tax benefit schedule.<sup>26</sup>

To see this, observe that the denominator of the accounting rate-of-return (whether total capitalization or stockholder's equity) will be the same before and after taxes. The numerator in the after-tax case, after-tax profits less depreciation, will be:

$$(A24)$$

$$\begin{aligned} & \int_{-\infty}^t (1-\alpha)(f(t-\theta) - d(t-\theta))I(\theta) d\theta, \\ &= \int_{-\infty}^t ((1-\alpha)f(t-\theta) + \alpha d(t-\theta))I(\theta) d\theta \\ & \quad - \int_{-\infty}^t d(t-\theta)I(\theta) d\theta, \\ &= \int_{-\infty}^t f^*(t-\theta)I(\theta) d\theta - \int_{-\infty}^t d(\theta)I(\theta) d\theta. \end{aligned}$$

But this is the *same* numerator as would be encountered in the before-tax analysis for a firm with the same depreciation schedule, but *before-tax* benefits  $f^*(\cdot)$ . For such a firm,  $r'$  would be the before-tax economic rate of return. Hence analysis of the after-tax case is

<sup>26</sup>A word about the treatment of inflation seems appropriate here. In the before-tax analysis it does not matter whether we work in real or nominal dollars (so long as we are consistent). In the after-tax case, however, the fact that depreciation which is deductible for tax purposes must be in nominal terms appears to raise some difficulty. That difficulty is only apparent however. Suppose that we begin by working in real terms. The nominal nature of the depreciation deduction plus the effects of inflation affect the depreciation schedule measured in real terms. We show, however, that any after-tax case with *any* depreciation schedule is isomorphic to a before-tax case. The effects being considered will, of course, influence *what* that before-tax case is, but they will not alter the existence of such a case. Hence, while the nominal nature of depreciation (like any other factor affecting the depreciation schedule) will affect what the numerical value of the real after-tax accounting rate of return is, it will not change our results.

identical to that of the before-tax case with an appropriate adjusted definition of the benefit schedule. All previous results apply to it.<sup>27</sup>

<sup>27</sup>It is interesting (and revealing of the full unity of the before- and after-tax analyses) to note what happens in the case of "economic depreciation" examined above. In that case, it turns out that the (pathologically unlikely) choice of an economic depreciation schedule involves the *same* depreciation schedule whether economic depreciation is chosen before or after tax. Assets are valued at the present value of all remaining benefits either before or after tax; it makes no difference. Further, that choice of depreciation schedule makes the after-tax economic rate of return  $r'$  relate to the before-tax economic rate of return  $r$ , in the natural (but—except with this depreciation schedule—not inevitable) way:  $r' = r(1-\alpha)$ . To show these things, return to the differential equation (equation (A5)) from which we derived the formula for economic depreciation in the before-tax case.

$$(a) \quad V'(\theta) = rV(\theta) - f(\theta).$$

Consideration of the before-tax analysis shows that if and only if  $V(\cdot)$  satisfies this and  $V(0) = 1$ , then

$$(b) \quad V(\theta) = \int_0^\infty f(u) \exp(-r(u-\theta)) du$$

the present value of future benefits. Now, choose  $V(\cdot)$  and hence  $d(\cdot)$  to satisfy (b) and therefore (a). Then

$$(c) \quad (1-\alpha)V'(\theta) = r(1-\alpha)V(\theta) - (1-\alpha)f(\theta),$$

$$(d) \quad V'(\theta) = r(1-\alpha)V(\theta)$$

$$-((1-\alpha)f(\theta) + \alpha d(\theta)) = r(1-\alpha)V(\theta) - f^*(\theta),$$

since  $d(\theta) \equiv -V'(\theta)$ . But this is in the same form as (a). Hence, as in (A6):

$$(e)$$

$$V(\theta) = \int_0^\infty f^*(u) \exp(-r'(u-\theta)) du + C \exp(r'\theta),$$

with  $r' \equiv r(1-\alpha)$ . Here,  $C$  is a constant of integration; however  $C=0$ , since (c) shows that  $V(\infty)$  is finite. Since  $V(0)=1$ , we have

$$(f) \quad 1 = \int_0^\infty f^*(u) \exp(-r'u) du,$$

which shows that  $r'$  is the after-tax economic rate of return. From (b) and (e) with  $C=0$ ,  $V(\theta)$  is both the before- and the after-tax present value of the remaining benefit stream at  $\theta$ , whence economic depreciation is the same in both cases. See Paul Samuelson.

Thus, in after-tax analysis as in before-tax analysis, there is no reason to believe that differences in the accounting rate of return correspond to differences in economic rates of return. Our computer examples show the effects can be very large; the belief that they are small enough in practice to make accounting rates useful for analytic purposes rests on nothing but wishful thinking.

## REFERENCES

- Adams, Walter, *The Structure of American Industry*, New York: Macmillan, 1970.
- Areeda, Philip and Turner, Donald F., *Antitrust Law: An Analysis of Antitrust Principles and Their Application*, Boston: Little, Brown and Co., 1978.
- Benston, George J., "The FTC's Line of Business Program: A Benefit-Cost Analysis," in Harvey Goldschmid, ed., *Business Disclosure: Government's Need to Know*, New York: McGraw-Hill, 1979, 58-118.
- Cooper, Joseph D., *Proceedings of the Second Seminar on Economics of Pharmaceutical Innovation*, Washington: American University, 1976.
- Fisher, Franklin M., "Diagnosing Monopoly," *Quarterly Review of Economics and Business*, Summer 1979, 19, 7-33.
- Hotelling, Harold, "A General Mathematical Theory of Depreciation," *Journal of the American Statistical Association*, September 1925, 20, 340-53.
- Livingstone, J. Leslie and Salamon, Gerald L., "Relationship Between the Accounting and the Internal Rate of Return Measures: A Synthesis and Analysis," in J. Leslie Livingstone and Thomas J. Burns, eds., *Income Theory and Rate of Return*, Columbus: Ohio State University, 1971, 161-78.
- Samuelson, Paul A., "Tax Deductibility of Economic Depreciation to Insure Investment Valuations," *Journal of Political Economy*, December 1964, 72, 604-06.
- Scherer, F. M., *Industrial Market Structure and Economic Performance*, Chicago: Rand McNally, 1980.
- Solomon, Ezra, "Alternative Rate of Return Concepts and Their Implications for Utility Regulation," *Bell Journal of Economics*, Spring 1970, 1, 65-81.
- Stauffer, Thomas R., "The Measurement of Corporate Rates of Return: A Generalized Formulation," *Bell Journal of Economics*, Autumn 1971, 2, 434-69.
- Weiss, Leonard W., "The Concentration-Profits Relationship and Antitrust," in Harvey J. Goldschmid et al., *Industrial Concentration: The New Learning*, Boston, Little, Brown and Co., 1974.
- U.S. Council of Economic Advisers, *Economic Report of the President*, Washington: USGPO, 1979.

# A Model of Housing Tenure Choice

By J. V. HENDERSON AND Y. M. IOANNIDES\*

What are the determinants of tenure choice in the housing market? In the empirical literature (see, for example, Harvey Rosen; David Laidler) at the top of the list for the United States are the income tax advantages of owning. Even given the excessive tax allowances for depreciation enjoyed by landlords, back-of-the-envelope calculations of equilibrium effective prices indicate that it is still financially advantageous to own rather than rent for all income brackets except for, say, the lowest-income quartile of families, unless a family is planning to move soon (see John Shelton). The qualification about moving enters because homeowners face higher transactions costs of selling and moving than renters.

While tax laws may favor owning in the United States, in many countries the tax treatment of renting vs. owning is financially unimportant, neutral, or biased towards renting. Therefore, in developing a general theoretical framework, it is desirable to first concentrate on the noninstitutional economic aspects of the problem, and then introduce consideration of country-specific institutional factors such as the tax system. This also helps us understand the impact of tax laws on tenure choice.

There are two basic approaches to the problem in the theoretical literature. One is to look at the factors affecting the choices of a single consumer. In looking at these factors, some recent writers concentrate on the risk avoidance behavior of consumers in particular situations facing uncertain future rents, housing prices, and rates of inflation (see, for example, John Bossons; Glenn Canner; Ioannides). Other writers concentrate on life cycle aspects of the problem.

Roland Artle and Pravin Varaiya examine a situation under perfect certainty where

people choose to own or rent based on how owning distorts their desired path of asset holdings. Owning may require purchases of assets at a time when dissaving would be preferred, such as for young people who are on a low part of their life cycle path of income, or for old people who want to consume their equity. Artle and Varaiya are able to attain fairly explicit life cycle solutions, but at the expense of two unusually restrictive assumptions. They assume everyone at a point in time and at every point over the life cycle must consume exactly the same quantity of housing, regardless of wealth, price, or temporary income differences. Also, it is assumed arbitrarily that the opportunity cost of owning is less than the cost of renting.

The other approach in the theoretical literature is to solve for market equilibrium. The only paper we know of which does this is by Yoram Weiss. While his study is interesting because there is a complete market solution dividing the world into renters and owners, the solution is derived under again unusually restrictive assumptions. In a world of perfect certainty, Weiss assumes that people are indexed by differing efficiencies in producing housing services from homes if they own-occupy. However, for some reason that efficiency is equal for all individuals when they rent. Given this dichotomy, not surprisingly those whose owner-efficiency parameter exceeds (is less than) the common rental efficiency parameter own (rent), in the absence of institutional factors (taxes).

In this paper, we propose to integrate the key economic elements underlying the papers in the literature and to relax the unusually restrictive assumptions. In particular, any differences in opportunity cost between renting and owning will be derived, housing consumption will be perfectly divisible and income and price responsive, and there will be uncertainty about future prices and returns. The cost of integration will be some loss in terms of explicitness of life cycle and

\*Brown University and Boston University, respectively. We thank an anonymous referee for comments which were helpful in writing the final draft.

market-equilibrium solutions. What we will be looking at are individuals operating in a housing market in equilibrium and seeing how their tenure choices are affected by wealth, life cycle, and other considerations. Given this analysis, we can then identify some of the critical characteristics of an equilibrium.

Before introducing the life cycle and portfolio considerations, we explicitly analyze the economic differences between the opportunity cost of renting and owning. This formally introduces a fundamental ingredient involved in tenure choice decisions for any durable, which as far as we know has been ignored in the theoretical literature. This is accomplished in Section I, where an externality associated with renting durables is identified and shown to be responsible for the relative attractiveness of owning.

Section II introduces uncertainty and deals with comparisons between owning and renting when housing may be held as an investment good in a portfolio of assets. The general proposition is that a person's housing consumption demand will not equal his housing investment demand. If a person's consumption demand for housing is less than the quantity of the housing stock desired for investment purposes, consumption demand can be accommodated by simply occupying part of one's investment stock. In the absence of capital market imperfections, people in this position will definitely owner-occupy the housing they consume. We show that the likelihood of a person being in this position is related critically to the time pattern of income receipts and the level of wealth.

If, on the other hand, in the portfolio problem, consumption demand for housing exceeds investment demand, a person may rent or owner-occupy. In analyzing this, we assume a person cannot split his tenure so that part of his consumption is owner-occupied and part is rented—a person cannot be in two places at the same time. Thus if he does owner-occupy in this situation, it means he must "distort" his investment and consumption choices to bring investment levels into equality with consumption levels. This "distortion" may be incurred voluntarily because the rental externality enhances

the attractiveness of owning relative to renting for individuals with income streams in an appropriately defined region.

Section III examines tax impacts and capital market imperfections.

### I. The Fundamental Rental Externality

There is a basic externality connected with the rental of a durable that, given equilibrium prices, makes it more attractive to own than to rent. To isolate this factor, we turn to analyzing the nature of durable good consumption in a world of perfect certainty. The services from a durable are a function of the capacity or stock, and of the rate of utilization (see, for example, Guillermo Calvo). Where  $h_c$  is capacity and  $u$  the rate of utilization, total services are

$$(1) \quad h = h_c f(u); \quad f' > 0, \quad f'' < 0.$$

Thus at the same rate of utilization, services double if we double capacity, but less than double if we double the rate of utilization for the same capacity.

The costs of greater rates of utilization include the resource costs of utilization itself such as time costs and the costs of increased maintenance and repairs. Events requiring maintenance or repairs might be broken into rough categories such as (i) breakdowns (for example, furnace or electrical system in a house), (ii) obvious damages (broken windows), and (iii) "hidden" wear and tear (chipped or dirtied paint on walls or scratched floors). All these costs can be summarized by the strictly convex function  $T(u)$  where total utilization costs are

$$(2) \quad h_c T(u); \quad T' > 0, \quad T'' > 0.$$

Nontime costs of utilization are incurred as maintenance required to restore the asset to its original condition. Alternatively, they could represent depreciation (capital loss) of the asset, or any combination involving partial restoration.

The externality involved in renting arises from the maintenance problem. An owner-occupier directly incurs  $h_c T(u)$ . But a landlord can collect from the tenant only part of

variations in  $h_c T(u)$ , over and above the basic contract rent. A tenant, of course, incurs his own time costs of utilization and, in addition, the tenant might be charged for items under the category of obvious damages through, for example, deductions from a damage deposit. However, the marginal costs of increased breakdowns and wear and tear caused by increased rates of utilization cannot be fully charged to the tenant. It is impossible to explicitly provide in rental contracts for all possible contingencies, let alone to even collect on all contingencies provided for. Thus tenants are assumed to pay less than owners at all rates of utilization.

Tenants pay

$$(3) \quad h_c \tau(u); \quad \tau' > 0; \quad \tau'' > 0;$$

where  $\tau(u) < T(u) \forall u$ ,  $\tau'(u) < T'(u) \forall u$ .

This presents a classic externality problem in the rental market, where tenants do not face the social marginal costs of their utilization rates. This is not the case for owner-occupiers and is a critical factor in the determination of the opportunity cost of renting vs. owning and occupying. We note it is also possible to rephrase this discussion in terms of a tenant being unable to collect from a landlord for improvements he makes on an apartment.

#### A. Renting vs. Owning

To isolate the impact of this externality, we examine a consumer deciding whether to rent or own under conditions of perfect certainty. Then we state the condition for market equilibrium and, based upon this condition, do a comparison of owning vs. renting. The following important assumptions are made:

(i) *The consumer maximizes a multiperiod utility function.* However, in terms of examining optimal period 1 decisions, we can bury a third and subsequent periods in the indirect utility function given remaining wealth at the beginning of the second period. This assumes that optimal decisions are made in future periods, and that for our comparative statics, future prices and incomes are

held constant. The same specification is also used later when uncertainty is incorporated.

(ii) *For housing consumed in period 1, costs of utilization are incurred in the second period only.*<sup>1</sup>

If he owns, the consumer's maximization problem is to choose the values  $\{x^*, h_c^*, S^*, u^*\}$  that maximize

$$(4) \quad U(x, f(u)h_c) + V(w),$$

subject to  $y_1 = x + Ph_c + S$

$$w = y_2 + S(1+r) + Ph_c - T(u)h_c.$$

The function  $U(\cdot)$  is the utility the consumer derives from the period 1 consumption bundle and  $V(\cdot)$  is the indirect utility function of wealth remaining after period 1. Both  $U(\cdot)$  and  $V(\cdot)$  are assumed to be increasing and strictly quasi concave. The term  $y_1$  is period 1 income;  $x$  is period 1 consumption of the numeraire, which is all other goods;  $h_c$  is housing stock purchased and  $u$  is its rate of utilization;  $S$  is period 1 savings which earns the market real rate of interest  $r$ ;  $P$  is the constant market purchase price of a unit of housing stock; and  $y_2$  is the income the consumer receives in the beginning of period 2. Price  $P$  can differ between periods without affecting the results (in equation (6) below,  $R$  adjusts accordingly). We shall use  $\nu^*$  to denote the maximum value of the consumer's lifetime utility, when he owns.

If he rents, the budget and wealth constraints of the consumer's maximization problem change. A renter, therefore, chooses the values  $\{\tilde{x}, \tilde{h}_c, \tilde{S}, \tilde{u}\}$  that maximize

$$(5) \quad U(x, f(u)h_c) + V(w),$$

subject to  $y_1 = x + S + Rh_c$ ;

$$w = y_2 + S(1+r) - \tau(u)h_c,$$

<sup>1</sup>While it might be more realistic to specify time or breakdown costs as being incurred in period 1, to conserve on notation and the number of functions we specify them as occurring only once, with no loss of generality.

where  $R$  is the rental price of housing. We shall use  $\bar{v}$  to denote the maximum value of lifetime utility when the consumer rents.

In market equilibrium, for there to be owners of rental housing, interest foregone on the equity in housing must equal housing profits. Thus for each unit of housing, equilibrium in asset holdings requires that

$$(6) \quad \frac{rP}{1+r} = R - \frac{(T(\bar{u}) - \tau(\bar{u}))}{1+r},$$

where  $\bar{u}$  indicates that this  $u$  is the tenant's and not his landlord's choice variable (although under perfect certainty the landlord must know what its value will be). Given the tenant's  $\bar{u}$ , the landlord directly recovers  $\tau(\bar{u})$  from the tenant, but must pay  $T(\bar{u})$  for utilization. Given the market-equilibrium purchase price of housing, for housing to be held as an asset by a landlord, rents are set at a level that covers the financial opportunity cost plus costs of utilization not directly recovered from the tenant, all appropriately discounted. Thus although tenants only directly pay  $\tau(\bar{u})$  for utilization, they indirectly pay the balance in the form of higher rents.

Finally in equation (6), we could specify savings for either renters or owner-occupiers as containing investment in housing. But given (6), no results would be altered by adding these terms on, since under perfect certainty from an investment point of view, investment in financial assets and investment in housing are perfect substitutes.

By maximizing utility for renters and owners and combining first-order conditions with equation (6), we can obtain the equilibrium rates of utilization under owning  $u^*$ , and renting  $\bar{u}$ , respectively. These are given by

$$(7) \quad \frac{f'(u^*)}{f(u^*)} (rP + T(u^*)) = T'(u^*),$$

$$(8) \quad \frac{f'(\bar{u})}{f(\bar{u})} (R(1+r) + \tau(\bar{u})) = \tau'(\bar{u}).$$

Note that the level of capacity chosen does not appear in (7) and (8) so that the equilibrium  $u^*$  and  $\bar{u}$  are independent of capacity

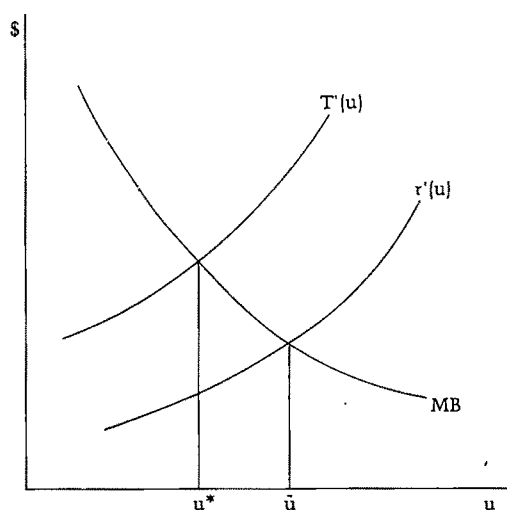


FIGURE 1

(this follows from the separability in (1) and (2)).

The left-hand side of these equations can be interpreted as the marginal benefits of increasing utilization. For an owner, the benefits are the increase in the rate of services from capacity,  $f'(u^*)$ , multiplied by the marginal utility of capacity (where  $(U_2/U_1)(1+r) = (rP + T(u^*)) \cdot f(u^*)^{-1}$ ). In equilibrium in a world of perfect certainty, the renter's  $\bar{u}$  would have to equal the  $\bar{u}$  of a landlord's tenant in equation (6). By substituting (6) into (8), we see that the equilibrium schedules of marginal benefits of increasing utilization as functions of  $u$  are the same for (7) and (8). They are graphed in Figure 1. The marginal benefit curve is downward sloping because  $f'' < 0$  (this can be seen by differentiating the left-hand side of (7) or (8)). The right-hand side of equations (7) and (8) are the marginal cost of increasing utilization where for any  $u$   $\tau' < T'$ , and  $\tau'', T'' > 0$ . The two curves are depicted in Figure 1. The obvious conclusion is that in equilibrium  $\bar{u} > u^*$ . The externality results in *overutilization* of capacity under rental tenure.

If there are no other considerations, renting is always inferior to owning. The variable  $u$  is chosen on the basis of  $\tau(u)$  rather than true costs. However, the tenant must indi-

rectly pay the balance in terms of higher contract rents (equation (6)). Formally to prove the desirability of owning, for any equilibrium  $P$ , we examine  $\bar{v}$ , defined as a Taylor-series expansion about  $v^*$ . Thus,

$$\begin{aligned}\bar{v} - v^* &= U_1^*(\bar{x} - x^*) + U_2^*(\tilde{h}_c f(\bar{u}) \\ &\quad - h_c^* f(u^*)) + V^*(\bar{w} - w^*) + G.\end{aligned}$$

For strictly quasi-concave functions,  $G < 0$ . Substituting in, using (6) where  $\bar{u} = \bar{u}$ , and rearranging yields

$$(9) \quad \frac{\bar{v} - v^*}{U_1^*} < \left\{ \frac{\delta}{f(u^*)} (rP + T(u^*)) - d \right\} \\ \times \tilde{h}_c (1+r)^{-1} < 0.$$

The variables  $\delta$  and  $d$  are defined by the Taylor-series expansions

$$f(\bar{u}) = f(u^*) + f'(u^*)(\bar{u} - u^*) + \delta$$

$$\text{and } T(\bar{u}) = T(u^*) + T'(u^*)(\bar{u} - u^*) + d,$$

where, by earlier assumptions on concavity and convexity,  $\delta < 0$  and  $d > 0$ . Owning dominates renting.

Note that owning dominates renting for everyone, even if people have differing efficiencies in their  $f(\cdot)$  functions. One might think that people who are less efficient in producing housing services would rent to try to shift the costs of high  $u$ 's to landlords. However, such a market equilibrium is not sustainable, even if landlords cannot discriminate among potential tenants on the basis of differing  $f(\cdot)$  functions. The proof is simple. Consider any potential group of people from the spectrum of  $f(\cdot)$  functions who might rent. For there to be any landlords, a version of equation (6) must be met where contract rents are sufficiently higher than expected maintenance costs to at least cover opportunity costs of investment. This means, however, that people in that group of potential renters with equal to or lower than expected maintenance costs will find it financially disadvantageous to rent, as in equation (9). Thus there can never be a group of

individuals from the spectrum of  $f(\cdot)$  functions who will rent in equilibrium.<sup>2</sup> For that to occur, we need some other reason for renting in the model, such as portfolio considerations or capital market imperfections.

## II. An Economic Basis for Renting

Durables such as housing may serve dual purposes for many consumers—as a consumption good and as an investment holding in a portfolio. To see this and its implications in a framework that has become a standard tool of such investigations (compare Agnar Sandmo, 1968), we examine the problem of a consumer simultaneously choosing his optimal housing consumption and his optimal portfolio.

We present the consumer-maximization problem in a two-period model. However, from the work of Paul Samuelson, Eugene Fama, and Hayne Leland, we also know this problem of optimal period 1 bundles can be embedded in an  $n$  period framework if we assume all investment and consumer decisions from the end of life back to period 1 are made optimally. The function that serves as the period 2 utility function contains as arguments price distributions in future periods, future incomes, and future discount factors. As long as these factors are held constant, we can derive propositions about the allocation of wealth among consumption and investment goods today. However, it should be noted for reference below that the magnitudes of relative and absolute risk-aversion measures are not insensitive to many of the factors which have been suppressed in representing future utility by the indirect utility of wealth remaining after period 1.

In the analysis to follow, we treat housing investment as a risky asset, relative to a safe financial asset. This conventional assumption may not be the most reasonable one.

<sup>2</sup>The referee suggested this situation could give rise to a signalling equilibrium where those with lower utilization rates would try to communicate this fact to the landlord. A signalling equilibrium would only be possible if lower utilization rates are negatively correlated with signalling costs (see A. Michael Spence). We do not see the basis for such an assumption in this case; also the topic is beyond the scope of the paper.

Because bonds and savings deposits in many countries (unlike Israel or Brazil) are not inflation-indexed, their real return may be more variable than the real return to investments in commodities such as housing (or soybeans). However, it should be noted that, for landlords, even if real capital gains have a low variance in a stable market, each landlord may face a high variance maintenance expenditure (unpaid by the tenant). Below we will explore the impact of altering the specification of which asset is safe versus risky.

*A. Renter vs. Owner-Occupier:  
Behavioral Models*

In the most general case the consumer chooses his consumption demand, investment holding of housing, savings, and the rate of utilization  $\{\tilde{h}_c, \tilde{h}_f, \tilde{S}, \tilde{u}\}$  so as to maximize

$$(10) \quad U(y_1 - S - (P - L - R)h_f - Rh_c, h_c f(u)) + E[V(y_2 + S(1+r) + (P(1+\theta) - L(1+r) - (T(\bar{u}) - \tau(\bar{u})))h_f - \tau(u)h_c)]$$

where only income is time-scripted;  $h_f$  is the consumer's investment in housing which he rents out to others in period 1 at a price of  $R$ . In period 2 he incurs noncollectable maintenance costs of  $(T(\bar{u}) - \tau(\bar{u}))h_f$ , that are uncertain given that the tenants chose  $\bar{u}$ . He can sell his asset for an unknown return of  $\theta$ . To avoid having to deal with the costs of defaults, we assume

$$P(1+\theta) - L(1+r) - (T(\bar{u}) - \tau(\bar{u})) > 0, \quad \forall \theta, \bar{u}.$$

The term  $L$  reflects the fact that he can obtain a mortgage loan of  $L$  at the fixed market rate of interest  $r$ . However, since (dis)savings,  $S$ , is an unconstrained choice variable for now, at the moment the presence of  $L$  is irrelevant. In Section III, however, it will play a critical role. Finally, the consumer

rents  $h_c$  units of housing capacity for consumption purposes at a price  $R$ , and later pays damage payments of  $\tau(\bar{u})h_c$ , given his preferred rate of utilization  $\bar{u}$ .

Maximizing (10) with respect to  $\{\tilde{h}_c, \tilde{h}_f, \tilde{S}, \tilde{u}\}$ , we get

$$(11a) \quad -U_1 R + U_2 f(\bar{u}) - E[V']\tau(\bar{u}) = 0;$$

$$(11b) \quad -(P - R - L)U_1 + E[V'(P(1+\theta) - L(1+r) - (T(\bar{u}) - \tau(\bar{u})))] = 0;$$

$$(11c) \quad -U_1 + E[V'](1+r) = 0;$$

$$(11d) \quad U_2 f'(\bar{u}) - E[V']\tau'(\bar{u}) = 0.$$

Let  $\bar{v}$  denote the maximum value of lifetime utility, (10).

For the optimal rate of utilization, we again obtain equation (8) by combining equations (11a), (11c), and (11d). Therefore,  $\bar{u}$  is independent of income, housing capacity, and portfolio considerations. Other straightforward manipulations yield

$$(12a) \quad \frac{U_2}{U_1} f(\bar{u}) = R + \frac{\tau(\bar{u})}{1+r};$$

$$(12b) \quad rP/(1+r) = R + \frac{E[V'(P\theta - (T(\bar{u}) - \tau(\bar{u})))]}{E[V'](1+r)}$$

Equations (12a) and (12b) have standard interpretations as marginal conditions determining  $\tilde{h}_c$  and  $\tilde{h}_f$ . In particular, (12b) expresses the standard asset equilibrium condition for assets with stochastic returns of  $P(\theta) - (T(\bar{u}) - \tau(\bar{u}))$ , assuming the tenant's  $\bar{u}$  is uncertain; the return to housing is adjusted above the return to the safe asset by the individual's risk margin.

From (12a) and (12b) it should be clear that in this general problem, optimal housing consumption demand and optimal investment demand are quite different. If we were facing a standard consumption-portfolio problem, our specification of the maximization problem would end here. We would turn

directly to the interpretation of the results (Part B), focusing on the issue of when consumption demand for housing exceeds or falls short of investment demand. But in equations (10)–(12), consumption and investment are divorced and there is no issue of tenure per se, because everyone rents their consumption. However, we made assumptions earlier about the nature of housing which will ensure that some people will want to owner-occupy their consumed housing.

If consumption demand is less than investment demand, for example, it would be efficient for the consumer to owner-occupy that part of his investment up to  $\tilde{h}_c$  and rent out the rest,  $\tilde{h}_I - \tilde{h}_c$ . In doing so, he will avoid the rental externality analyzed in Section I; and thus it can be shown that he will be better off (Section III). If consumption demand exceeds investment demand, however, we assume he cannot own only part of his consumption; that is, his consumption tenure cannot be “split.” But to avoid the rental externality, if  $\tilde{h}_c$  is near  $\tilde{h}_I$ , he could “distort” his investment and consumption choices, raise  $h_I$  to  $h_c$ , and owner-occupy his entire investment.<sup>3</sup>

Because consumption tenure cannot be split and because of the rental externality, the maximization problem in (10) in fact only applies to people who are actually going to rent their consumed housing. Renters are taken from the group of people for whom  $\tilde{h}_c > \tilde{h}_I$ . All people for whom  $\tilde{h}_c \leq \tilde{h}_I$  and some for whom  $\tilde{h}_c > \tilde{h}_I$  will own. We turn to their maximization problem now.

### 1. The Owner-Occupier's Problem

An owner-occupier's lifetime utility maximization problem is to choose the values  $\{h_c^*, h_I^*, S^*, u^*\}$  that maximize

$$(13) \quad U(y_1 - S - (P - L)h_I + R(h_I - h_c), h_c f(u^*)) + E\{V(y_2 + S(1+r) - T(u)h_c - (h_I - h_c)(T(\bar{u}) - \tau(\bar{u})) + [P(1+\theta) - L(1+r)]h_I)\},$$

subject to the constraint

$$(13') \quad h_I - h_c \geq 0.$$

Constraint (13') is indispensable for the validity of the expression (13). Secondly in (13), for own-consumed investment, the consumer gives up  $Rh_c$  in rental income in period 1 and pays utilization costs of  $T(u)h_c$  in period 2. On rented-out investment, he pays uncompensated utilization costs of  $(h_I - h_c)(T(\bar{u}) - \tau(\bar{u}))$  in period 2.

If we let  $\mu$  be the Lagrange multiplier corresponding to constraint (13'), the optimal consumption bundle  $\{h_c^*, h_I^*, S^*, u^*\}$  must satisfy the following conditions:

$$(14a) \quad -RU_1 + f(u^*)U_2 + E[V'(T(\bar{u}) - \tau(\bar{u}) - T(u^*))] = \mu;$$

$$(14b) \quad (P - L - R)U_1 - E[V'(P(1+\theta) - L(1+r) - (T(\bar{u}) - \tau(\bar{u})))] = \mu$$

$$(14c) \quad -U_1 + (1+r)E[V'] = 0;$$

$$(14d) \quad f'(u^*)U_2 - T'(u^*)E[V'] = 0;$$

$$(14e) \quad \mu = 0, \quad h_I^* - h_c^* \geq 0;$$

$$(14f) \quad \mu > 0, \quad h_I^* - h_c^* = 0.$$

Equation (14a) indicates that the owner-occupier by avoiding the rental externality “gains back”  $T(\bar{u}) - \tau(\bar{u})$  per unit of housing investment used for own consumption.

If  $h_I^* > h_c^*$  so  $\mu = 0$ , the marginal condition for determining  $h_I^*$  is the same as (12b). However, the other marginal conditions are different. For  $h_c^*$  and  $u^*$  we have

$$(15a) \quad U_2 f(u^*)/U_1 = \left( R - \frac{E[V'(T(\bar{u}) - \tau(\bar{u}))]}{E[V'](1+r)} \right) + \frac{T(u^*)}{1+r};$$

$$(15b) \quad \frac{f'(u^*)}{f(u^*)} \left( R(1+r) + T(u^*) - \frac{E[V'(T(\bar{u}) - \tau(\bar{u}))]}{E[V']} \right) = T'(u^*).$$

<sup>3</sup>This is tantamount to a one-directional indivisibility.

The opportunity cost of consuming an additional unit of one's housing investment is the value of net rent foregone (the first two terms on the right-hand side of (15a), plus the maintenance cost associated with owner-occupancy). Equation (15b) is similar to (8). However, if  $\bar{u}$  is uncertain, the optimal  $u^*$  now is no longer independent of income and housing consumption, since the level of consumption chosen affects the *uncertain* utilization rebate ( $T(\bar{u}) - \tau(\bar{u})$ ) from increased own-consumption and owner-occupancy utilization costs, thus affecting risk considerations.

If  $\mu > 0$  and  $h_I = h_c$ , then the interpretation of the first-order conditions is modified accordingly. Combining (14a) and (14b), we get

$$(16a) \quad f(u^*)U_2/U_1 = (P - L) - E[(V'/U_1) \\ \times (P(1 + \theta) - L(1 + r) - T(u^*))].$$

Substituting in (14c) into (16a) yields

$$(16b) \quad f(u^*)U_2/U_1 = P(r/(1 + r)) \\ - E[V'\theta]/U_1 + T(u^*)/(1 + r).$$

Housing consumption, which is equal to investment, is adjusted until the marginal evaluation of an additional unit of capacity equals its marginal cost. Its marginal costs are marginal utilization costs,  $T(u^*)/(1 + r)$ , plus the value of expected opportunity costs (foregone interest,  $rP/(1 + r)$ ), less the value of expected capital gains  $PE[V'\theta]/U_1$ .

#### B. Who Owner-Occupies and Who Rents

Let us turn to the central question of this section. What types of people are likely to rent vs. own their consumed housing; that is, for whom will  $\tilde{h}_I < \tilde{h}_c$  vs.  $\tilde{h}_I^* \geq \tilde{h}_c^*$ ? Using the above models, we can specify conditions under which the impact on tenure choice of differences in total wealth and differences in the slope of the path, or tilt, of the income stream can be unambiguously stated.

Total wealth consists of two components: initial income or wealth,  $y_1$ , and second-period income  $y_2$ . Total wealth increases if

and only if  $dy_1 + dy_2/(1 + r) > 0$ . The time path of income is tilted if  $dy_1 - dy_2/(1 + r) \neq 0$ . If  $dy_1 - dy_2/(1 + r) > 0$ , the time path is tilted towards period 1. If  $dy_1 - dy_2/(1 + r) < 0$ , it is tilted towards period 2.

In addition to these definitions, we make one critical assumption. The wealth elasticity of demand for housing consumption (in the unconstrained maximization problem (10)) is positive; that is,  $dh_c/d(y_1 + y_2/(1 + r)) > 0$ . In short, we assume housing consumption is not an inferior good.

The basic results can be derived by examining the general consumption-portfolio problem in (10). In doing comparative statics we assume  $U$  in (10) is additively separable, we differentiate equations (11a, b, c) with respect to  $y_1$  and  $y_2$ , and solve for changes in  $h_I$ ,  $h_c$ , and  $S$ . Note (11d), or its equivalent (8), is not differentiated because  $\bar{u}$  is independent of  $y_1$ ,  $y_2$ , and all preference parameters. By differentiating and solving we get

$$(17a) \quad dh_I = -\frac{U_{11}U_{22}}{D}(1 + r)(f(\bar{u}))^2 \\ \times E[(\beta + \gamma - \xi)V''][(1 + r)dy_1 - dy_2];$$

$$(17b) \quad dh_c = \frac{U_{11}}{D}(R(1 + r)^2 + (1 + r)\tau) \\ \times \{E[V'']E[(\beta + \gamma)^2V''] \\ - (E[(\beta + \gamma)V''])^2\} \left( dy_1 + \frac{1}{1 + r}dy_2 \right).$$

Also we note

$$(17c) \quad dS = \{ \tau(1 + r)^{-1}U_{11}(R(1 + r)^2 \\ + \tau(1 + r)) (E[V''(\beta + \gamma)^2]E[V''] \\ - E[V''(\beta + \gamma)]^2) + U_{11}U_{22}f(\bar{u})^2 \\ \times E[V''(\beta + \gamma)(\beta + \gamma - \xi)] \} D^{-1} \\ \times (dy_1 - dy_2/(1 + r)).$$

The auxiliary variables are defined as

$$\begin{aligned}\beta &\equiv P - L(1+r); \\ \gamma &\equiv P\theta - (T(\bar{u}) - \tau(\bar{u})); \\ \alpha &\equiv y_2 - \tau(\bar{u})h_c + S(1+r); \\ \xi &\equiv (P - L - R)(1+r).\end{aligned}$$

The determinant  $D$  of the Hessian obtained in differentiating equations (11a, b, c) is given by

$$\begin{aligned}D &= U_{11}U_{22}(f(\bar{u}))^2 E[(\beta + \gamma - \xi)^2 V''] \\ &+ \{U_{11}(R(1+r) + \tau(\bar{u}))^2 \\ &+ U_{22}(f(\bar{u}))^2(1+r)^2\} \\ &\times \{E[V''']E[(\beta + \gamma)^2 V''] \\ &- (E[(\beta + \gamma)V'''])^2\}.\end{aligned}$$

For a consumer to be at a maximum, the Hessian must be negative definite and a necessary condition for this is that  $D < 0$ .

To interpret the results in equations (17), we utilize common concepts in the theory of risk aversion. We use the coefficient of the degree of absolute risk aversion ( $A$ ) where  $A = -V'''/V'$ . Kenneth Arrow and subsequent writers argue that it is most "reasonable" to assume that  $A$  declines as wealth rises, or that there is decreasing absolute risk aversion. Both cases are considered, although we will focus on the case of decreasing  $A$ . For later reference, we will also define the relative risk aversion coefficient  $F = -wV'''/V'$ .

Examining equations (17), we can prove the following. (Proofs where not obvious are in the Appendix.)

(i) If wealth rises ( $dy_1(1+r) + dy_2 > 0$ ) with no change in the tilt of the path of income ( $dy_1(1+r) - dy_2 = 0$ ),  $dh^I = 0$ ; or the demand for portfolio holdings of housing does not change. Similarly the demand for safe assets is unchanged.

(ii) If and only if  $V$  exhibits decreasing risk aversion ( $dA/dw < 0$ ), the portfolio de-

mand for housing increases as income is tilted towards period 1. That is,  $dh^I/d(y_1(1+r) - y_2) > 0$  iff  $dA/dw < 0$ . Correspondingly, the portfolio demand for housing declines as income is tilted towards the future.

(iii) The consumption demand for housing, holding wealth constant, is unaffected by the tilt of income. By assumption  $dh_c$  increases as wealth increases. (As an aside, we note a sufficient but not necessary condition for this assumption to hold is that  $V$  exhibits both decreasing absolute risk aversion ( $dA/dw < 0$ ) and increasing relative risk aversion ( $dF/dw > 0$ ). These are also sufficient conditions for  $D$  to be negative given concavity of  $U$  and  $V$ .)

Given these results, the relative attractiveness of owner-occupancy vs. renting for individuals who have identical preferences but differ with respect to period 1 and 2 incomes can be demonstrated. We do so graphically in Figure 2 by examining regions of the  $(y_1, y_2)$  plane where people rent vs. owner-occupy. By result (iii), differences in  $y_1$  and  $y_2$  which leave total wealth unchanged leave housing consumption demand  $h_c$  unchanged. Thus, the loci of points on the  $(y_1, y_2)$  plane for which  $h_c$  is constant are straight lines with slope  $-(1+r)$ . Assuming an increasing wealth demand for housing,  $h_c$  increases in the northeast direction as marked by  $h_c^1, h_c^2, h_c^3, \dots$ .

By result (i), the loci of points on the  $(y_1, y_2)$  plane for which the portfolio holdings of housing are the same are straight lines with slope  $(1+r)$ , along which  $dy_1 = dy_2(1+r)^{-1}$ . For wealth constant, a tilt from  $y_2$  to  $y_1$  increases investment demand, if and only if  $dA/dw < 0$  from result (ii). Thus, demand for  $h_I$  increases in the southeast direction as marked by the lines  $h_I^1, h_I^2, h_I^3$  if and only if absolute risk aversion is decreasing. If  $dA/dw > 0$ ,  $h_I$  decreases in the southeast direction.

The curve  $\Delta\Delta$  illustrates a possible locus of points for which  $\hat{h}_I = \hat{h}_c$  as  $y_1$  and  $y_2$  change. Under the assumption of decreasing absolute risk aversion,  $\Delta\Delta$  must have slope less than  $(1+r)$  and greater than  $-(1+r)$ , so both  $h_c$  and  $h_I$  move together. (If there is increasing absolute risk aversion,  $\Delta\Delta$  would have a slope greater than  $(1+r)$  or less than  $-(1+r)$ .)

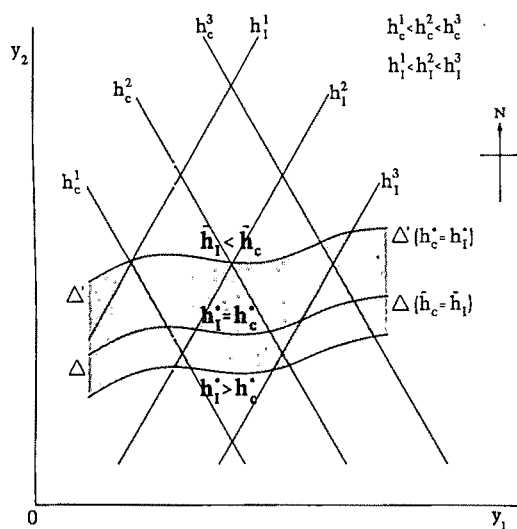


FIGURE 2

The  $\Delta\Delta$  locus divides the plane into regions in a general consumption portfolio where consumption demand exceeds or falls short of investment demand. In our specific application to housing, people to the south of  $\Delta\Delta$  would owner-occupy because investment demand exceeds consumption demand. For those to the north of  $\Delta\Delta$ , their consumption demand *potentially* exceeds their investment demand; and they are potential renters (see subsection 3 for further analysis).

### 1. Wealth and Tenure Choice

The people to the north of  $\Delta\Delta$ , for a given  $y_1(1+r)/y_2$ , are those who have higher wealth, as we move up any  $h_1$  curve. Thus, for any relative time path of income, *higher wealth people will be renters*. This result does not depend on the nature of risk aversion. With increasing absolute risk aversion, an untilted increase in wealth increases consumption demand relative to portfolio demand. In that case, those to the east of the relevant  $\Delta\Delta$  (not shown) would be renters; again these are high-wealth people.

Similarly the result does not depend on the characterization of housing as being the relatively risky asset. The demand for housing consumption also rises relative to the safe asset ( $S$ ) with an untilted increase in wealth.

The observed patterns of tenure choice are at odds with this completely general result. The question is, why? Part of the explanation may be that in correlating wealth and tenure choice empirically, life cycle considerations may not be controlled for, so that many of the assumed low-wealth people are those with strongly tilted income streams (see below). Even so, the presumption is that controlling for tilt, empirically, owner-occupancy rates increase with wealth. The explanation in our model must lie in factors other than unconstrained portfolio-consumption demands, such as the rental externality, tax laws, and capital market imperfections. We explore each of these in turn below.

### 2. Income Path and Tenure Choice

In Figure 2, it should be apparent that at a given level of wealth there may be both renters and owners. Assuming  $dA/dw < 0$ , holding wealth constant, *renters are those with lower  $y_1(1+r)/y_2$ 's, or those whose income streams are tilted towards the future*. This would imply, for example, that among the young, renters would tend to be the more educated because their income streams tend to be more tilted to the future. Also, those with more inherited financial wealth relative to human capital wealth will have income streams tilted to the present, and will tend to own. We do not view these as unexpected results. Thus our real concern is accounting for the observed wealth effects on tenure choice.

### 3. Impact of the Rental Externality

*Intuitive Analysis.* The demarcations in Figure 2 using  $\Delta\Delta$  are based on problem (10) which in fact is specific only to people who actually rent their consumption. We know from Part A of this section that owner-occupiers face a revised maximization problem; and this would apply to those who would initially desire  $\tilde{h}_1 \geq \tilde{h}_c$ , because of the maintenance externality. Also those for whom  $\tilde{h}_1 < \tilde{h}_c$  but  $\tilde{h}_1 \rightarrow \tilde{h}_c$  are probably better off "distorting" their choices, equating  $h_1$  and  $h_c$ , owner-occupying, and avoiding the rental externality. This suggests in our illustrative solution that  $\Delta\Delta$  is only a hypothetical locus and that the actual boundary

between renters and owners is another curve, say  $\Delta'\Delta'$ , which lies above  $\Delta\Delta$ . In essence, the rental externality effectively shifts the boundary between renters and owners in favor of owning.

In fact, what we expect is a region to the north of  $\Delta'\Delta'$ , where everyone rents, a strip on the plane (with  $\Delta'\Delta'$  being the north boundary) where there is the constraint  $h_I^* = h_c^*$  in equations (14) and (16), and the region south of the strip where  $h_I^* > h_c^*$  and people hold housing investments over and above their own homes. This strip is drawn as stretching below  $\Delta\Delta$  for reasons cited below in the discussion of Figure 3.

The rental externality thus affects the relationship between wealth and tenure choice, in the sense that it stretches out the region of wealth in which people owner-occupy. However, unless the externality is so costly as to eliminate renting, we would still expect regions where, controlling for tilt, the highest wealth people would still rent their consumption. This does *not* imply that low-wealth people would be the direct landlords of high-wealth people, especially since the holdings of a single low-wealth person might not provide the housing for a single high-wealth person. Rather low- and high-wealth people (who also have asset holdings of housing) would pool their money through stock holdings in housing corporations. These corporations would be the direct landlords of high-wealth tenants.

*Technical Analysis.* To understand the impact of the rental externality on  $\Delta\Delta$  in Figure 2, we turn to Figure 3 where we plot indirect utility against period 2 income with period 1 income being held constant. We also assume decreasing absolute risk aversion, so that renters are those with low  $y_1/y_2$ 's. Indirect utility is increasing in total wealth. The solid curve  $\tilde{v}$  that turns into a dashed one at  $m_1$  as  $y_2$  declines, plots the maximal value of utility for the renter's problem, equation (10). The solid part, as  $\tilde{v}$  increases beyond  $m_1$ , denotes equilibrium utility for renters. The solid curve  $m_5m_1$  plots the equilibrium values of  $v^*$ , the maximal value of utility under owner-occupancy. The move from  $m_1m_2$  to  $m_1m_5$  can be broken into two parts. The plotting of  $\tilde{v}$  has a discontinuity at  $m_2$ , where it jumps to the curve  $m_3m_4$ . At  $m_2$ ,  $\tilde{h}_c = \tilde{h}_I$ ;

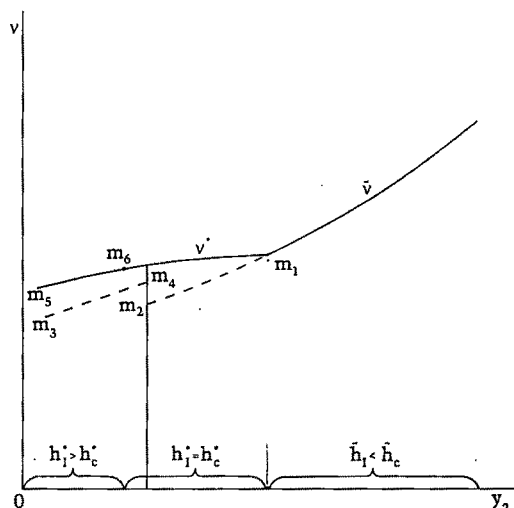


FIGURE 3

The jump comes because once  $\tilde{h}_c \leq \tilde{h}_I$ , for any  $\tilde{h}_c, \tilde{h}_I$  combination, the consumer could owner-occupy  $\tilde{h}_c$  of his  $\tilde{h}_I$  investment and avoid the rental externality. However, once owner-occupancy occurs, the correct maximization problem is given in (13) and this switch to the correct maximization problem again raises utility to  $m_1m_5$ . Along  $m_5m_1$ , from  $m_5$  to, say,  $m_6$   $h_I^* > h_c^*$ , at  $m_6$   $h_I^* = h_c^*$  where  $\mu = 0$  in (14), and from  $m_6$  to  $m_1$   $\mu > 0$  so that  $h_I^* = h_c^*$  by constraint.

It is not clear in Figure 3 whether  $m_6$ , where  $h_c^* = h_I^*$ , occurs to the left or right of  $m_2$ , where  $\tilde{h}_c = \tilde{h}_I$ . We draw it to the left based upon the following reasoning. Shifting to owning at  $m_2$  for the same  $h_c$  and  $h_I$  would raise "disposable" wealth because of the savings on the externality. Given the same tilt to the income path, this wealth effect would raise  $h_c$  relative to  $h_I$  so, at that  $y_2$ , we would expect  $h_c^* > h_I^*$ , although  $\tilde{h}_c = \tilde{h}_I$ . If  $h_c^* > h_I^*$  then  $h_c^* = h_I^*$  to the left of  $m_2$ . For this same reason, the  $\Delta'\Delta'$  strip in Figure 2 extends south of  $\Delta\Delta$ .

Thus along the dashed extension of  $\tilde{v}$ ,  $m_1m_2$  where  $\tilde{h}_I < \tilde{h}_c$ , it pays people to distort their choices and set  $h_I = h_c$  solving problem (14) for the case  $\mu > 0$ . The tradeoffs involved in adopting this distortion can be seen by doing a Taylor-series expansion of  $\tilde{v}$  about  $v^*$ .

A Taylor-series expansion reveals three factors in the tradeoff made by consumers as

to whether to owner-occupy or rent.<sup>4</sup> The first two are obvious. The first is the owner-occupancy savings from avoiding the maintenance externality already documented in equation (9). The second is the owner-occupancy portfolio distortion cost of changing  $h_r$  to  $h_c$  which is the increase in investment multiplied by the marginal cost of the increased risk incurred over and above the market risk premium given. The third factor is less obvious. A tenant's choice of the utilization rate,  $\bar{u}$ , may differ from the utilization rate anticipated by landlords,  $E[\bar{u}]$ . If  $\bar{u} > E[\bar{u}]$ , there is an additional benefit to renting because a portion of the marginal cost of the renter's excessive maintenance rate is not borne by him. Thus, for example, those whose  $f(\cdot)$  functions are shaped such as to result in unusually high  $\bar{u}$ 's will find it advantageous to rent. The opposite is the case for those whose functions are shaped such that  $\bar{u} < E[\bar{u}]$ .

### III. Other Considerations

#### A. Taxes

Suppose individuals face income taxes. Imputed rents to owner-occupiers are tax exempt while rental income of landlords is

<sup>4</sup>We compare utility from renting with utility from owning when  $h^* = h^* = h$ . We do a Taylor-series expansion of  $\bar{v}$  about  $v^*$  and substitute in for  $\bar{x}$ ,  $x^*$ ,  $\bar{w}$ , and  $w^*$  from equations (10) and (13), from (14), and from the Taylor-series expansions of  $f$  and  $f^*$  and  $T$  and  $T^*$  from (9). Details are available from the authors upon request. Rearranging the result gives

$$\begin{aligned} \bar{v} - v^*/U_1 &< (\delta_j f^* U_2^*/U_1^* - E[V'/U_1^*] d) \bar{h}_c \\ &- \bar{h}_c E[(V'/U_1^*)((T(\bar{u}) - \tau(\bar{u}) - (T(\bar{u}) - \tau(\bar{u}))))] \\ &+ (\bar{h}_c - \bar{h}_r)(-cov[V'/U_1^*, P\theta \\ &- T(\bar{u}) - \tau(\bar{u})] - Prem), \end{aligned}$$

where *Prem* is the market-equilibrium risk premium on housing. When choosing to own, and hence bring  $h_c$  and  $h_r$  into equality,  $-cov(\cdot)$  is the premium required to fully compensate the individual for incurring the increased risk alone (ignoring the other factors). *Prem* is the smaller premium available in the market given the equilibrium of demand and supply of housing as an investment good.

taxable. The tax treatment of interest payments is the same for rental and owner-occupied units. In short, the tax system makes owner-occupying cheaper than renting. Assume momentarily that the income tax rate is constant so that individual tax burdens are proportional. In this case, the impact of the tax system on tenure choice is identical to that of the rental externality (after removing the uncertainty facing landlords about the rate of utilization  $\bar{u}$ ). Owner-occupancy per unit of housing is cheaper than renting by some constant proportion. In Figure 2, this simply stretches out the region of wealth in which people owner-occupy. As with the rental externality, unless the tax advantage is so great as to eliminate renting entirely, we would still find a region, where controlling for the tilt of the income path, high-wealth people rent their consumption.

An effectively progressive tax system, however, could eliminate the region of high-wealth renters, implying everyone would owner-occupy their consumption in the absence of transactions costs of moving. The reason is that, although the discrepancy between desired consumption and investment increases with wealth, so does the per unit subsidy to owner-occupied consumption. There will be a set of pairs of average tax rates and marginal tax rates (degree of progressivity) above which renting would be eliminated.

#### B. Capital Market Imperfections

Capital market imperfections can be introduced into our model by imposing the constraint in all maximization problems that  $S \geq 0$ . That is, consumers cannot borrow against future income for current consumption. However, they can borrow to purchase durables which can be offered as collateral. Thus the mortgage loan term  $L$  in the previous sections now has a meaningful role.

Formally, the constraint  $S \geq 0$  changes all previous first-order conditions (equations (11c) and (14c)) on savings from  $-U_1 + (1+r)E[V'] = 0$  to

$$\begin{aligned} (18) \quad &-U_1 + (1+r)E[V'] = 0, \quad \text{if } S^* \leq 0; \\ &-U_1 + (1+r)E[V'] \leq 0, \quad \text{if } S^* = 0. \end{aligned}$$

All our analyses would require adjustment to account for how the shadow price  $\gamma$  of the constraint ( $S \geq 0$ ) changes as total wealth and income stream tilt change. Under *relevant mortgage loan terms*, the principal impact of the constraint is to make it less attractive to own, since owning requires risky investment, when in fact, either total dissavings is desired to increase current consumption or dissavings in the safe asset is desired (but prohibited) in order to finance purchases of the risky asset. This can be seen by using a Taylor-series expansion to compare renting and owning as we did in Section II.

Relevant mortgage loan terms are that, per unit of housing,  $P - L - R > 0$ ; or  $P - L > R$  so that an owner-occupier requires a greater gross cash outlay to consume a unit of housing than a renter. Thus renting is attractive since current consumption of all other goods is higher than under owning, as is desired when  $\gamma > 0$ . Why the constraint that  $P - L - R > 0$ ? If  $P - L - R < 0$ , for any potential landlord, the net outlays on housing investments are negative, indicating that everyone will have an infinite demand for housing investments unless default is costly. Collateral has little meaning if negative net outlays are allowed.

We consider two issues with capital market imperfections. What is their impact on tenure choice and how do people who are constrained by the imperfections behave? The basic wealth effects on tenure choice are unchanged by capital market imperfections, in a model in which housing consumption rises with wealth, and loans cover some maximizing fraction of investment. For high-wealth people, if the tilt of their income is such that  $S \geq 0$  is a binding constraint, then raising their investment towards their consumption demand to owner-occupy is even more costly than it was before. Capital market imperfections would only increase the size of the region in Figure 2 where high-wealth people rent. It would also enlarge the region of lower-wealth people with incomes tilted to the future who rent. But, in summary, this standard specification of capital market imperfections is not sufficient to alter our basic types of findings.

We could, of course, eliminate low-wealth owners by introducing a capital market im-

perfection where a minimum wealth level was required to obtain a loan, or there was a minimum size requirement on a mortgage (eliminating mortgages for cheaper homes). If there are sufficient fixed costs to making loans (irrespective of loan size), then a rational banker could impose such a restriction.

How do people behave who are constrained by the imperfections? We examine only the most interesting case, which applies to those who are already constrained in owner-occupancy to have  $h^* = h^c$ . We want to know what are the impact of income and of loan term changes on the demand for housing. For example, suppose  $y_2$  increases for  $y_1$  constant. The resulting increase in total wealth enhances housing consumption demand. However, the tilting of income to the future lowers total portfolio demand and increases the desire to dissave.

For this problem, assuming that the constraints  $S = 0$  and  $h^I = h^c$  remain binding, the relevant set of necessary conditions are one for housing consumption (which equals investment), that is equation (16a), and equation (14d) for  $u^*$ . Differentiating these equations with respect to  $h^*$ ,  $u^*$ ,  $y_1$ , and  $y_2$ , for  $dh^*$  we obtain

$$(19) \quad dh^* = \frac{dy_1}{D} \left[ (P - L) U_{11} (f'' U_2 + h(f')^2 U_{22} - T'' E[V'] + h(T')^2 E[V'']) \right] \\ - \frac{dy_2}{D} \left\{ E[V''(\beta + \gamma)] (f'' U_2 + h(f')^2 U_{22} - T'' E[V']) + T' E[V''] (h f' U_{22}) \right\},$$

where we have used auxiliary variables defined earlier;  $D$  is the determinant of the Hessian obtained in differentiating (16b) and (14d); and  $D > 0$  for a maximum.

Concavity of  $U$ ,  $V$ , and  $f$  and convexity of  $T$  alone ensure that the expression in the square brackets of the  $dy_1$  term is positive. An increase in  $y_1$  increases  $h$  which is expected since the consumption (wealth) and housing investment effects (increased portfolio demand where only housing is in the portfolio) move together. These same condi-

tions also ensure that an increase in  $y_2$  decreases  $h$ . Therefore, the increase in  $y_2$  causes a reduction in total portfolio demand dominating any other possible effects. This reduction can only be met by reducing housing investment since  $S = 0$  already.

Finally, we note that the impact of changing loan terms,  $L$ , can be analyzed in the same fashion. By differentiating (16b) and (14d) with respect to  $L$ , we get the rather obvious result that an increase in loan size per unit of housing increases the demand for housing. Consumption and investment demands go together in this case. The loan reduces the cost of the constraint of no dissavings, and housing consumption demand is enhanced since more money for general consumption is available in period 1. While this result is obvious, some applications are less so.

The analysis of a change in  $L$  demonstrates the impact of expected inflation on housing demand. Expected inflation under the traditional U.S. mortgage contract where payments are a nominal fixed stream tilts the housing payment stream towards the purchase period. That is, the real value of the down payment and initial mortgage payments rise relative to more distant payments. That, in effect, is formally equivalent in our model to a reduction in the loan size, raising the initial payment and lowering the later one. Thus, people operating under our constraints ( $S = 0$ ,  $h_c^* = h_f^*$ ) will experience reduced demand for housing under the traditional mortgage contract. Robert Schwab devotes much of his 1982 paper to showing just this.

#### IV. Conclusions

A model designed to illuminate the dual role of housing as a consumption and investment good was successful in giving a whole host of interesting predictions about housing market behavior and the determinants of tenure choice. Some of these predictions agree with stylized facts and others can be tested empirically.

In our model, housing stock is used to produce housing services and as an investment good. The amount of housing services produced per unit of housing stock depends

upon the rate of utilization, which is assumed to be chosen by occupants, whether renters or owners. Equilibrium in the holding of assets implies that owning housing stock does not differ from holding any other asset unless uncertain rates of return are introduced. With perfect certainty, tenure choice was shown to depend on an important externality associated with renting. We showed that the presence of the rental externality along with equilibrium in asset holding implies that owning dominates renting. In addition, we showed that the equilibrium rate of utilization of housing stock for renters exceeds that of owners, and both are independent of all individual characteristics. They depend only on market prices, technological characteristics, and maintenance charge schedules.

Renting becomes more attractive if housing is subject to random capital gains or losses and consumers may also invest in a capital market at a fixed rate of return. In the model with uncertainty, the advantage which the rental externality confers on owner-occupancy has to be weighed against the characteristics of consumers' risk avoidance behavior. We showed that individuals who either have less wealth or who receive relatively more of their wealth in the beginning of their lifetimes will be the net lenders of housing, and vice versa for renters. These results hold under the set of restrictions on preferences which ensure that the consumption demand for housing increases with total wealth. In the course of developing these results, we showed that within certain income ranges, because of the maintenance externality, it pays people to "distort" their investment and consumption choices and owner-occupy rather than rent.

Finally we examined the impact of tax laws favorable to owner-occupancy and of capital market imperfections on tenure choice. Apart from sufficiently large benefits to owner-occupancy from either the rental externality or average levels of taxation, which eliminate renting entirely, sufficient progressivity of the income tax structure was the only assumption which eliminated the region of high wealth renters. Capital market imperfections inhibited owner-occupancy, but did not alter the general pattern of re-

sults. Finally, we showed that, with capital market imperfections when consumption and investment demand for housing are constrained to be equal, housing consumption increases with first-period income, decreases with second-period income, and increases with the proportion of housing cost that can be mortgaged.

#### APPENDIX

We shall prove that under the assumptions  $dA/dw < 0$  and  $dF/dw \geq 0$ , the term

$$E[V''']E\{(\beta + \gamma)^2 V'''\} - (E\{(\beta + \gamma)V'''\})^2$$

is positive. In the course of doing this proof, the signs of all relevant parts of equations (17) and the determinant  $D$  will be explored. We add and subtract the term  $E[V''']E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\}$  to obtain

$$\begin{aligned} & E[V''']E\{(\beta + \gamma)^2 V'''\} - (E\{(\beta + \gamma)V'''\})^2 \\ & + E[V''']E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\} \\ & - E[V''']E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\} \\ & = E[V''']E\{(\beta + \gamma)(\beta + \gamma - \beta - \gamma + \xi)\} \\ & + E[V''']E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\} \\ & - (E\{(\beta + \gamma)V'''\})^2 \\ & = E\{(\beta + \gamma)V'''\}[\xi E[V'''] \\ & - E\{(\beta + \gamma)V'''\}] \\ & + E[V''']E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\} \\ & = E[V''']E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\} \\ & \quad (-) \quad (-) \\ & - E\{(\beta + \gamma)V'''\}E\{(\beta + \gamma - \xi)V'''\} \\ & \quad (-) \quad (+) \end{aligned}$$

By definition we have that  $V''' < 0$ ,  $\beta + \gamma > 0$ .

We shall now examine the signs of the terms:

$$E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\},$$

$$E\{(\beta + \gamma - \xi)V'''\}.$$

For the first term, by multiplying by  $h_I$  and by adding and subtracting  $\alpha E\{(\beta + \gamma - \xi)V'''\}$ ,

$$\begin{aligned} & E\{(\beta + \gamma)(\beta + \gamma - \xi)V'''\}h_I \\ & + E\{(\beta + \gamma - \xi)V'''\}\alpha \\ & - E\{(\beta + \gamma - \xi)V'''\}\alpha \\ & = E\{(\beta + \gamma - \xi)V'''\} [h_I(\beta + \gamma) + \alpha] \\ & + \alpha E\{(\xi - (\beta + \gamma))V'''\} \\ & = E\{(\beta + \gamma - \xi)V'''\}w \\ & + \alpha E\{(\xi - (\beta + \gamma))V'''\}. \end{aligned}$$

If we can prove both of these expressions are negative, the proof is complete. By using the definitions of absolute and relative risk-aversion coefficients, we can write the above expressions as equal to

$$\begin{aligned} & -E\{(\beta + \gamma - \xi)V'F\} \\ & - \alpha E\{(\xi - (\beta + \gamma))V'A\}. \end{aligned}$$

If we define  $\bar{A}$  and  $\bar{F}$  as the values of  $A$  and  $F$  when the state of nature  $\theta$  is such that  $\beta + \gamma - \xi = 0$  then the above becomes:

$$\begin{aligned} & E\{(\bar{F} - F)(\beta + \gamma - \xi)V'\} \\ & + \alpha E\{(\bar{A} - A)(\xi - (\beta + \gamma))V'\}, \end{aligned}$$

given from (11b) and (11c)  $E\{(\beta + \gamma - \xi)V'\} = 0$ ; where  $\beta + \gamma - \xi > [ < ] 0$ , we have  $\bar{A} - A > [ < ] 0$ , if  $A$  is decreasing with  $w$ , and  $\bar{A} - A < [ > ] 0$ , if  $A$  is increasing with  $w$ . Clearly  $E\{(\beta + \gamma - \xi)V'''\} = 0$ , iff  $A$  is constant and is positive [negative] iff  $A$  is decreasing [increasing] with  $w$ . Similarly, where  $\beta + \gamma - \xi > [ < ] 0$ ,  $\bar{F} - F < [ > ] 0$ , if  $F$  is

nondecreasing with  $w$ . As a result, the first of the above terms is negative, and so is the second, if  $\alpha > 0$ , and  $F$  is assumed to be nondecreasing and  $A$  is assumed to be decreasing with  $w$ .

## REFERENCES

- Arrow, Kenneth J., *Aspects of the Theory of Risk Bearing*, Helsinki: 1965.
- Artle, Roland and Varaiya, Pravin, "Life Cycle Consumption and Ownership," *Journal of Economic Theory*, June 1978, 18, 35-58.
- Bossoms, John D., "Housing Demand and Household Wealth," in L. S. Bourne and J. R. Hitchcock, eds., *Urban Housing Markets*, Toronto: University of Toronto Press, 1978.
- Calvo, Guillermo A., "Efficient and Optimal Utilization of Capital Service," *American Economic Review*, March 1975, 65, 181-86.
- Canner, Glenn, "Redlining and Mortgage Lending Patterns," in J. V. Henderson, ed., *Research in Urban Economics*, Vol. 1, Greenwich: JAI Press, 1981.
- Fama, Eugene F., "Multiperiod Consumption-Investment Decisions," *American Economic Review*, March 1970, 60, 163-74.
- Ioannides, Yannis M., "Temporal Risks and the Tenure Decision in Housing Markets," *Economic Letters*, 1979, 4, 293-97.
- Laidler, David, "Income Tax Incentives for Owner-Occupied Housing," in A. C. Harberger and M. J. Bailey, eds. *The Taxation of Income from Capital*, Washington: Brookings Institution, 1969.
- Leland, Hayne E., "Saving and Uncertainty: The Precautionary Demand for Saving," *Quarterly Journal of Economics*, August 1968, 82, 465-73.
- , "Optimal Growth in a Stochastic Environment," *Review of Economic Studies*, January 1974, 41, 75-86.
- Rosen, Harvey S., "Housing Decisions and the U.S. Income Tax: An Econometric Analysis," *Journal of Public Economics*, February 1979, 11, 1-24.
- Samuelson, Paul A., "Lifetime Portfolio Decisions by Stochastic Dynamic Programming," *Review of Economics and Statistics*, August 1968, 51, 239-46.
- Sandmo, Agnar, "Portfolio Choice in a Theory of Saving," *Swedish Journal of Economics*, June 1968, 70, 106-22.
- , "The Effect of Uncertainty of Saving Decisions," *Review of Economics Studies*, July 1970, 37, 353-60.
- Schwab, Robert M., "Inflation Expectations and the Demand for Housing," *American Economic Review*, March 1982, 72, 143-53.
- Shelton, John P., "The Cost of Renting vs. Owning a Home," *Land Economics*, February 1968, 44, 59-72.
- Spence, A. Michael, *Market Signaling*, Cambridge: Harvard University Press, 1974.
- Weiss, Yoram, "Capital Gains, Discriminatory Taxes, and the Choice Between Renting and Owning a House," *Journal of Public Economics*, August 1978, 10, 45-55.

# On Divergence of Opinion and Imperfections in Capital Markets

By JORAM MAYSHAR\*

The importance of divergence of opinion in the functioning of capital markets was recognized by early economic writers. In the prevailing models of capital markets, however, differences of opinion either do not exist or do not matter. Thus, although heterogeneity of opinion is allowed in the models developed by Kenneth Arrow, Gerard Debreu, and Peter Diamond, nothing essential would change if all individuals were to hold identical, homogeneous-equivalent average expectations.

In the capital asset pricing model (CAPM) of William Sharpe (1964) and John Lintner (1965), homogeneous expectations are assumed at the outset. When they considered the implications of heterogeneity of expectations in the model, both Lintner (1969) and Sharpe (1970) reached similar conclusions; as stated by Sharpe, "in a somewhat superficial sense the equilibrium relationships derived for a world of complete agreement can be said to apply to a world in which there is disagreement, if certain values are considered to be averages" (p. 291). Sharpe's conclusion was that "a model based on disagreement has little value in a positive role" (p. 113).

The aim of this paper is to present a simple model of exchange in capital markets where divergence of opinion not only exists, but is essential. It is essential because of its association with endogenous limitations on the number of active market participants. It will be argued that in the models cited above, the significance of divergence of opinion was dismissed because of the failure to recognize the implications of the obvious fact that investors choose not only the size of their holdings in each asset, but also in which

assets to invest. Correspondingly, they failed to recognize that in (imperfect) capital markets, equilibrium requires the simultaneous determination of asset prices and the identity of investors trading in each asset.

As both Lintner (1969) and Sharpe (1970) recognized, the case of divergent opinions may differ from the case in which there is no such divergence, if only because it implies that investors may seek to sell short assets that they believe to be overrated. Lintner, who pursued the implications of the case in which short sales are not allowed, argued that, when not all investors trade in every asset, the price of an asset will reflect an *average* of the assessments of only those investors who *actually* hold the asset. Lintner, however, did not realize that, given that the set of active investors is endogenously determined, this is an incomplete characterization of how asset prices are determined. It leaves an integral question unanswered; namely, what distinguishes the active from the nonactive investor? Lintner was thus led to dismiss an alternative characterization of equilibrium asset prices—the marginal-investor theory—proposed by John Maynard Keynes and John Burr Williams thirty years earlier.<sup>1</sup>

Keynes characterized the determination of asset prices in a manner that attests to the importance that he attached to divergence of opinion among investors: "The prices of capital assets move until...they offer an

\*The Hebrew University of Jerusalem and The Maurice Falk Institute for Economic Research. I am indebted to and thank Gerald O. Bierwag, Charles Manski, and Shlomo Yitzhaki for comments and suggestions on earlier drafts.

<sup>1</sup>Lintner stated: "Any carryover of...Ricardian notations of 'marginal' buyers setting prices in purely competitive markets is *utterly unjustified and misleading* when dealing with security markets under uncertainty. Every investor is a *marginal* holder with respect to his last share...of *each* security he holds" (1969, p. 371). An early explicit statement of the proposition that it is the opinion of marginal traders which determines market prices was presented by Eugen von Bohm-Bawerk. A recent restatement of the marginal-investor theory was presented by Edward Miller.

equal apparent advantage to the *marginal investor* who is wavering between one kind of investment and another" (p. 217, emphasis added). This characterization was elaborated and formalized by Williams in his seminal work on *The Theory of Investment Value*. In his model, investors are presumed to formulate a single point estimate of asset values; each investor then invests his entire funds in the assets with the highest excess of (subjectively assessed) value over price. The key issue for investors, thus, is not how much to invest, but rather which asset to invest in. While each investor has a step demand curve, the market demand curve slopes down because of the heterogeneity of investors, and the market price is determined by the marginal opinion.<sup>2</sup>

The concern of Keynes and Williams with the endogenous determination of portfolio composition led them to the marginal-opinion theory. Lintner and Sharpe, on the other hand, who were concerned with the size of holdings and ignored the composition issue entirely, were led to the average-opinion theory. What I aim to demonstrate in this paper is that these seemingly conflicting theories are *both* correct; they are in fact complementary.<sup>3</sup> Asset prices are here shown to depend on the opinion of both average and marginal investors. The structure of opinion thus be-

comes essential in the determination of market equilibrium.<sup>4</sup>

The model presented in this paper can be considered as a simple extension of the *CAPM*. It adopts the same framework: a static one-good, two-period exchange model in which relevant trading occurs only once, in the "present," before the uncertainty of the future clears up. The model in fact imposes a stricter structure than the *CAPM* on investors' preferences and on the probability distribution of assets. It generalizes the *CAPM* by incorporating trading costs and heterogeneity of opinion among investors.

Transaction costs were examined earlier in my 1979 article. As discussed there, such costs may to a large extent substitute for the informational gaps between market participants. (This, in particular, seems to be the nature of the substantial costs, in the form of collateral requirements and their like, on short sales.) In that paper, however, investors were assumed to share the same expectations. Thus, while the number of assets in portfolios was determined, the actual composition of these only partially diversified portfolios was not. The heterogeneity of investors considered here solves this indeterminacy, and facilitates the description of how investors choose the composition of their portfolios.<sup>5</sup> Moreover, transaction costs on

<sup>2</sup>An identical model was employed by James Tobin to illustrate Keynes' theory of how liquidity preference determines the interest rate. (Williams had provided the same model two decades earlier.) As Tobin states in this context, "when [Keynes] refers to uncertainty in the market, he appears to mean disagreement among investors...rather than subjective doubt in the mind of an individual investor" (p. 70). In contrast to Keynes' use of the first interpretation of uncertainty as the basis for his liquidity preference theory, Tobin suggested using the second interpretation for that purpose.

<sup>3</sup>It is interesting to note that when Williams considered the possibility that investors would supplement their funds by borrowing, he anticipated the general nature of the results obtained here. He recognized that in this case, risk considerations would limit the scale of borrowing and thus endogenize the scale of investment, and that, as a result, "market price will...still be determined by marginal opinion concerning ultimate value, but the location of the margin will be affected by further opinions on the part of certain intramarginal owners..." (p. 22).

<sup>4</sup>Even though the subsequent discussion refers only the Sharpe-Lintner model, I believe that the general argument applies to the models of Arrow, Debreu, and Diamond as well. The basic issue is to recognize the economic significance of the Kuhn-Tucker conditions for the endogenous separation between active traders and self-excluded nontraders. (This issue has long been recognized in the partial equilibrium literature on the industry supply curve, and is nowadays also recognized in the literature on labor supply.)

<sup>5</sup>Heterogeneity of opinion together with transaction costs and short-sales restrictions were also recently analyzed by Robert Jarrow and in my 1981 article. Both works, however, follow Lintner (1969) and do not determine endogenously the composition of investors' partially diversified portfolios. Thus they do not provide a complete and explicit characterization of the equilibrium prices. The focus of my 1981 analysis is to demonstrate that the own-risk of an asset may have a much greater impact on its price than is recognized by the *CAPM*. The discussion of this important issue will not be repeated here.

asset purchases do not have here the essential role they played in that earlier paper. Results of the general nature claimed here can be obtained when opinions are sufficiently divergent and short sales are restricted even when there are no purchase costs.

The plan of the paper is as follows. The basic model is presented in Section I, the individual investor's optimization is analyzed in Section II, and market equilibrium conditions are derived in Section III. The implications of the model concerning the effects of diversity of opinion on equilibrium prices and concerning the informational content of prices are then examined in Section IV. It must be stressed at the outset that the aim of this paper is to illustrate propositions on the effects of divergence of opinion in the presence of market imperfections (rather than to provide a general characterization of the equilibrium). The somewhat long discussion of the individual investor's problem in Section II and the various simplifications introduced along the way should be understood with that aim in mind.

### I. The Model

In a standard two-period model, let there be  $N$  potential investors,  $M$  risky assets, and a single, zero-indexed, safe asset. The uncertain future wealth of investor  $i$  takes the form

$$(1) \quad y_i = q_{i0}(1+r) + \sum_{j \in T_i} q_{ij}x_j,$$

where  $q_{ij}$  denotes the quantity of asset  $j$  held by investor  $i$  and  $T_i$  is the set of risky assets in which he trades (i.e.,  $q_{ij} = 0$  for all  $j$  between 1 and  $M$  which are not in  $T_i$ ). The uncertain payout of a unit of asset  $j$  is denoted by  $x_j$ , and  $r$  is the safe interest rate, which is assumed to be exogenously determined.

Each investor is assumed to have the simplest form of mean-variance preferences:

$$(2) \quad U_i(y_i) = E_i y_i - (A_i/2) \text{Var}_i y_i.$$

Investors thus have a constant measure of

(absolute) risk aversion  $A_i$ ,<sup>6</sup> and, as indicated by the subscripts on the expectation operators, they use their own subjective probability assessments. Investors' expectations are considered here as exogenously determined.<sup>7</sup>

It is further assumed that asset payouts are distributed according to a diagonal (single index) structure,

$$(3) \quad x_j = \bar{x}_j + b_j z + e_j.$$

Thus the correlation between assets is only through the common risk factor  $z$  (representing the business cycle). The remaining stochastic factor  $e_j$  represents the specific risks of asset  $j$  so that

$$\text{Cov}_i(z, e_j) = \text{Cov}_i(e_j, e_k) = \text{Var}_i \bar{x}_j = 0$$

for all  $j$ , and all  $k \neq j$ . Investors may hold different opinions on the means and variances of assets; they are assumed, however, to share the same assessments of the covariance parameters  $b_j$ .

Up to now the model is but a special case of the CAPM which allows for heterogeneous expectations. Where the model now departs from the CAPM is in the specification of investors' budget constraints. It is assumed that individuals start in a cash position  $w_i$  and trade each asset separately. Trade in all assets except the safe asset is assumed to involve both fixed and proportional transaction costs.<sup>8</sup> Since the costs of purchasing an asset (when  $q_{ij} > 0$ ) are different in nature from the costs of short sales (when  $q_{ij} < 0$ ),

<sup>6</sup>This assumption is equivalent to assuming expected utility maximization with constant absolute risk aversion and normally distributed future wealth. It simplifies the subsequent analysis considerably since, together with the assumption on the existence of a riskless asset, it implies that all wealth effects are absorbed in the demand for the riskless asset. (Note, however, that  $A_i$  may well be inversely related to initial wealth.)

<sup>7</sup>Since the purpose here is to characterize the market equilibrium, rather than to explain how it is reached, the equilibrium prices can be included in the information sets on which investors base their expectations. The issue of what information may be recovered from the equilibrium prices is discussed in Section IV.B below.

<sup>8</sup>Transaction costs on trade in the safe asset and transaction costs (and taxes) on the future sale of assets are examined in my 1981 paper.

the trade set  $T_i$  has to be partitioned into two—the sets  $L_i$  and  $S_i$  of (long-) purchased and short-sold assets, respectively.

The budget constraint thus takes the form:

$$(4) \quad q_{i0} + \sum_{j \in L_i} [c'_{ij} + q_{ij}p_j(1 + t'_{ij})] + \sum_{j \in S_i} [c^s_{ij} + q_{ij}p_j(1 - t^s_{ij})] = w_i.$$

Here  $c'_{ij}$  and  $c^s_{ij}$  represent the fixed transaction costs (which are independent of the volume of trade) for a purchase or a short-sale of asset  $j$  by investor  $i$ . The corresponding proportional transaction costs (which are deducted from the proceeds of a short sale) are denoted by  $t'_{ij}$  and  $t^s_{ij}$ . The various transaction costs depend on the index  $i$  since they may include subjective elements such as the value of the time and effort devoted to the processing of trade-related information. The transaction costs on short sales may be particularly high, since they would have to include the costs of providing guarantees against default.

Two major implicit assumptions behind (4) deserve comment. First, it should be recognized that for all market participants, the net purchase price exceeds the net price of a sale. Thus, the model assumes away the existence of financial intermediaries whose superior handling of transaction costs may create a reverse net-price differential, which would make it profitable for them to engage in the simultaneous purchase and sale of the same assets.

A second important assumption in (4) is that all investors trade in their *entire* portfolio. This assumes away yet another commonly recognized implication of transaction costs concerning the desirability of a partial revision of the portfolio.<sup>9</sup> The initial funds  $w_i$  can thus be interpreted as the proceeds of a previously held portfolio which was entirely revised, or as labor income (possibly, in an

overlapping generation model, where the sellers of assets are the investors of an older generation). In either interpretation, the market supply of assets is assumed to be exogenous.

The problem faced by the individual investor in this model is to select disjoint sets  $L_i$  and  $S_i$  (their union being  $T_i$ ) and appropriately signed quantities  $q_{ij}$  for assets in these sets, in order to maximize the utility (2), given (1) and (4). This optimization is analyzed in the next section. Once we have optimized for all investors and determined the composition of each one's portfolio, we have also determined the sets,  $N_j$ , of the active investors who trade in each asset  $j$  (that is, those for whom  $j \in T_i$ ).

The equilibrium prices of assets would then be such that for each risky asset  $j$ , the total demand  $\sum_{i \in N_j} q_{ij}$  equals the exogenously given supply  $Q_j$ . This equilibrium condition, discussed in Section III, is different from and more general than the one in the CAPM in that the sets  $N_j$  will in general depend on market prices, rather than being exogenously given.

In order to simplify the exposition, it is assumed from here on that transaction costs on short sales are prohibitive, so that only purchases have to be considered. The sets  $L_i$  and  $T_i$  thus coincide (the more general case with short sales will be dealt with in footnotes).

## II. The Investor's Portfolio Problem

### A. The General Case

By substituting for  $q_{i0}$  from the budget constraint (4), the investor's utility (in the absence of short sales) can be expressed as

$$(5) \quad U_i = w_i(1 + r) + \sum_{j \in T_i} (q_{ij}\pi_{ij} - c_{ij}) - \frac{1}{2}A_i \text{Var}_i \left( \sum_{j \in T_i} q_{ij}x_j \right).$$

Here,  $c_{ij} = c'_{ij}(1 + r)$ ,  $t_{ij} = t'_{ij}(1 + r)$ , and  $\pi_{ij}$ , representing investor  $i$ 's subjective risk pre-

<sup>9</sup>Without this assumption the problem at hand would become hopelessly complicated. Equilibrium prices would be dependent not only on the complex structure of expectations and transaction costs, but on the structure of initial asset ownership as well.

mium for asset  $j$ , is defined by

$$(6) \quad \pi_{ij} = E_i x_j - p_j(1 + r + t_{ij}).$$

The investor's problem is then to select  $T_i$  and  $q_{ij} \geq 0$  for  $j \in T_i$  in order to maximize (5).<sup>10</sup>

This maximization problem can be divided into two stages. First, for any given  $T_i$ , optimize with respect to  $q_{ij}$ , and obtain the maximum utility  $U_i^*(T_i)$ ; then select the set  $T_i$  to maximize  $U_i^*(T_i)$ . For the first stage, in which a quadratic function is maximized over a closed linear set, a unique solution is guaranteed by standard calculus methods. (Since  $T_i$  is given at this stage, the fixed costs are sunk and do not affect the optimization.) For the second stage, which involves a discrete optimization, a solution is also guaranteed (since there exists only a finite number of different sets  $T_i$ ), although it may not be unique.

Given any portfolio composition  $T_i$ , the Kuhn-Tucker conditions for the first-stage optimization with respect to  $q_{ij}$  are

$$(7) \quad \pi_{ij} - A_i \text{Cov}_i(x_j, y_i) + \lambda_j = 0 \quad j \in T_i,$$

where  $\lambda_j \geq 0$ ,  $q_{ij} \geq 0$ , and  $\lambda_j q_{ij} = 0$  for all  $j$ . Thus  $\lambda_j = 0$  whenever the constraint  $q_{ij} \geq 0$  is not binding. Given the existence of fixed transaction costs, inspection of (5) suggests that the investor would not select a set  $T$  for which  $q_{ik} = 0$  for some  $k \in T$  is optimal, since if asset  $k$  were withdrawn from this set, a higher utility (by  $c_{ik}$ ) would be obtained. Thus, we may confine our attention already at this first stage of the optimization to those sets  $T_i$  where an internal solution, with  $q_{ij} > 0$  and  $\lambda_j = 0$ , holds for all assets.

Assuming that  $T_i$  is such a set, we can rewrite (7), using the diagonal structure (3), as

$$(8) \quad \pi_{ij} = A_i \left[ q_{ij} \text{Var}_i e_j + b_j \left( \sum_{k \in T_i} b_k q_{ik} \right) \text{Var}_i z \right]; j \in T_i.$$

<sup>10</sup>The weak inequality constraint is used to guarantee a solution even for nonoptimal sets  $T_i$ . As will be argued presently, the stricter restriction  $q_{ij} > 0$  is part of the solution and need not be considered as part of the problem faced by the investor.

As can easily be verified, this system of linear equations has the explicit solution

$$(9) \quad q_{ij} = \frac{1}{A_i \text{Var}_i e_j} [\pi_{ij} - b_j \pi_{iz}(T_i)]; j \in T_i,$$

where  $\pi_{iz}(T_i)$  which can be identified as  $i$ 's subjective premium for the risk  $z$ , is defined by

$$(10) \quad \pi_{iz}(T_i) = \frac{\sum_{k \in T_i} \pi_{ik} b_k (\text{Var}_i z / \text{Var}_i e_k)}{1 + \sum_{k \in T_i} b_k^2 (\text{Var}_i z / \text{Var}_i e_k)}.$$

The condition

$$(11) \quad \pi_{ij} > b_j \pi_{iz}(T_i) \quad \text{all } j \in T_i,$$

then guarantees that  $q_{ij} > 0$  does indeed hold for all assets  $j$  in the set  $T_i$ . Assuming that  $T_i$  satisfies condition (11), we may substitute (9) into (5) (noting that by (7),  $\sum_{j \in T_i} q_{ij} \pi_{ij} = A_i \text{Var}_i y_i$ ) to obtain the maximum utility,

$$(12) \quad U_i^*(T_i) = w_i(1 + r) + \sum_{j \in T_i} \frac{1}{2 A_i \text{Var}_i e_j} [\pi_{ij}^2 - b_j \pi_{ij} \pi_{iz}(T_i) - \phi_{ij}^2].$$

Here another key parameter is introduced, representing investor  $i$ 's purchase-threshold premium for the specific risk of asset  $j$ ,

$$(13) \quad \phi_{ij} = (2 A_i c_{ij} \text{Var}_i e_j)^{1/2}.$$

The second-stage optimization problem left for the investor is now to select the set  $T_i$  from all sets satisfying (11), in order to maximize (12). One case where it is easy to determine whether a particular asset  $k$  ought to be included in the optimum set  $T_i$  is when that asset is uncorrelated to all others, that is, when  $b_k = 0$ . In this case, an inspection of (12) reveals that a necessary condition (which also satisfies (11)) for the asset to be purchased is simply

$$(14) \quad \pi_{ik} \geq \phi_{ik}.$$

Strict inequality in (14) is a sufficient condition for trade in this uncorrelated asset  $k$ , since it is then clear from (12) that asset  $k$

increases utility, regardless of the composition of  $T_i$ . Equality in (14) indicates that this asset is marginal in the sense that the investor will be indifferent between trading in it (at the amount  $\phi_{ik}/A_i \text{Var}_i e_j$ ) and refraining from trade.<sup>11</sup>

Using the definition of  $\pi_{iz}(T_i)$  in (10),  $U_i^*(T_i)$  of (12) can be expressed in the alternative form

$$(15) \quad U_i^*(T_i) = w_i(1+r) + \sum_{j \in T_i} \frac{1}{2A_i \text{Var}_i e_j} \{ [\pi_{ij} - b_j \pi_{iz}(T_i)]^2 - \phi_{ij}^2 \} + \pi_{iz}^2(T_i)/2A_i \text{Var}_i z.$$

Using this form it seems tempting to attempt to generalize criterion (14) by requiring that  $j \in T_i$  whenever

$$(16) \quad \pi_{ij} - b_j \pi_{iz}(T_i) > \phi_{ij}.$$

This condition, however, neglects the effect which the inclusion of asset  $j$  in the trade set  $T_i$  may have on the factor  $\pi_{iz}(T_i)$ . It is thus not a correct criterion.<sup>12</sup> Even if corrected, such a criterion for the inclusion or exclusion of a *single* asset may not be sufficient, since combinations of assets will also have to be considered.<sup>13</sup>

<sup>11</sup>When short sales are allowed, the analysis can be extended with minor changes. Define  $\pi_{ij}^s = E_i x_j - p_j(1+r-t_{ij}^s)$  and  $\phi_{ij}^s = (2A_i c_{ij}^s \text{Var}_i e_j)^{1/2}$ , where  $t_{ij}^s = t_{ij}^s(1+r)$  and  $c_{ij}^s = c_{ij}^s(1+r)$ . Equations (12), (15), and (10) can then be left intact if we interpret summations such as  $\sum_{j \in T_i} \pi_{ij}$  as a short-hand notation for  $\sum_{j \in L_i} \pi_{ij} + \sum_{j \in S_i} \pi_{ij}^s$ , (and similarly for  $\phi_{ij}$ ). Both  $\pi_{iz}$  in (10) and  $U_i$  in (12) and (15) will then depend on  $L_i$  and  $S_i$  rather than on their union  $T_i$ . The necessary condition for short selling an uncorrelated asset  $k$  will be  $\pi_{ik}^s < -\phi_{ik}^s$ , and this will also guarantee that in such case  $q_{ik} < 0$ .

<sup>12</sup>A correct criterion, which may be obtained by somewhat tedious manipulations of (12), is that utility is increased by adding asset  $j$  to a given set  $T$  (when  $j \notin T$ ), if  $\pi_{ij} - b_j \pi_{iz}(T) > \phi_{ij}[1 + b_j^2(\text{Var}_i z / \text{Var}_i e_j)/\beta_i(T)]$ , where  $\beta_i(T)$  is the denominator of  $\pi_{iz}(T)$  in (10). The criterion for withdrawing asset  $j$  from a given  $T$  (when  $j \in T$ ) can be obtained by reversing the inequality and changing the plus sign on the right-hand side to a minus sign.

<sup>13</sup>To illustrate how complex this could be, consider a simple example with three assets. Omitting the index  $i$ , assume  $b_j = 1$  and  $\text{Var}_i e_j = \text{Var}_i z$  for all  $j$ ;  $\pi_1 = 0.1$ ,  $\pi_2 = 0.09$ ,  $\pi_3 = 0.08$ ,  $\phi_1 = 0.02$ ,  $\phi_2 = 0.03$ , and  $\phi_3 = 0.02$  (the

The source of these complications is the dependence of the factor  $\pi_{iz}$  on the composition of  $T_i$ . As a result, an individual's decision on whether to invest in a given asset  $j$  may depend on his expectations about all other assets. In turn, the equilibrium price of asset  $j$  may therefore also depend on the expectations of *all* investors concerning *all* assets.

A drastic simplification which cuts through much of this complexity, yet which nevertheless preserves the essential features of the problem at hand, is to modify the model by allowing individuals to trade costlessly among themselves in the risk asset  $z$ . The rest of this paper will employ this modified model.<sup>14</sup>

### B. The Modified Case

When investors can trade costlessly in the common risk factor  $z$ , the form of both future wealth (1) and budget constraint (4) will change. If  $q'_{iz}$  is the quantity of  $z$  traded by investor  $i$ , then  $q'_{iz}z$  will have to be added to his future wealth  $y_i$  in (1), and  $q'_{iz}p_z$  will have to be added on the left-hand side of the budget constraint (4). Equivalently we can consider the *total* holding of the  $z$  factor,  $q_{iz} = q'_{iz} + \sum_{j \in T_i} q_{ij}b_j$  to be the investor's decision variable instead of the net additions  $q'_{iz}$ . In this case, after substituting for  $q_{i0}$  and using the diagonal structure (3), the investor's utility function, corresponding to

discussion of an alternative example in my 1979 paper indicates that these figures are in a reasonable range). From (10) we get  $\pi_z(1) = 0.05$ ,  $\pi_z(2) = 0.045$ ,  $\pi_z(3) = 0.04$ ,  $\pi_z(1,2) = 0.0633$ ,  $\pi_z(1,3) = 0.06$ ,  $\pi_z(2,3) = 0.0567$ , and  $\pi_z(1,2,3) = 0.0675$ . A proxy  $\hat{U}$  for the maximum utility (12) can be obtained as the sum of the terms in the square brackets in (12) multiplied by  $10^4$ . Thus,  $\hat{U}(1) = 46$ ,  $\hat{U}(2) = 31.5$ ,  $\hat{U}(3) = 28$ ,  $\hat{U}(1,2) = 47.67$ ,  $\hat{U}(1,3) = 48$ ,  $\hat{U}(2,3) = 35.67$ , and  $\hat{U}(1,2,3) = 45.75$ . Consider the combination (1,2): we can see that there is nothing to be gained from withdrawing either of the assets (even though (16) is violated for asset 2) or from adding the third. Nevertheless, the best set is (1,3). Furthermore, even though asset 2 is superior to asset 3 in the absence of asset 1, in conjunction with asset 1 this ranking is reversed.

<sup>14</sup>A similar modification was used in my 1979 paper. It should be noted that results similar to those obtained in Section III below can also be obtained in the original unmodified model, by using alternative simplifications which restrict the form of heterogeneity of investors.

(5), will become

$$(17) \quad U_i = w_i(1+r) + \sum_{j \in T_i} [q_{ij}(\pi_{ij} - b_j\pi_{iz}) - c_{ij}] + q_{iz}\pi_{iz} - \frac{1}{2}A_i \text{Var}_i \left( \sum_{j \in T_i} q_{ij}e_j + q_{iz}z \right),$$

where  $\pi_{iz} = E_{iz} - p_z(1+r)$ .

Optimization with respect to  $q_{ij}$  results in first-order conditions similar to (9) where  $\pi_{iz}(T_i)$  is replaced by  $\pi_{iz}$ ,

$$(18) \quad q_{ij} = \frac{\pi_{ij} - b_j\pi_{iz}}{A_i \text{Var}_i e_j} \quad \text{for } j \in T_i;$$

$$q_{iz} = \frac{\pi_{iz}}{A_i \text{Var}_i z}.$$

To guarantee  $q_{ij} > 0$ , it is sufficient here to require  $\pi_{ij} > b_j\pi_{iz}$ . The maximum utility  $U_i^*(T_i)$  which is generated by a set  $T_i$  for which  $\pi_{ij} > b_j\pi_{iz}$  for all  $j \in T_i$  is, likewise, the utility given in (15), with  $\pi_{iz}(T_i)$  replaced by  $\pi_{iz}$ .

In the modified model, criterion (16) becomes applicable. A sufficient condition for asset  $j$  to be in the optimum portfolio is

$$(19) \quad \pi_{ij} - b_j\pi_{iz} > \phi_{ij}.$$

An important feature of criterion (19) is that the decision whether or not to invest in asset  $j$  is independent of both the prices of other assets and the expectations concerning those other assets. Equality in (19) implies that the investor is indifferent with regard to trading in asset  $j$ . In that case, we can say that asset  $j$  is a *marginal asset* for investor  $i$ , or, alternatively, that investor  $i$  is a *marginal investor* in asset  $j$ .

### III. Market Equilibrium

Conditions (18) and (19) can now be used to characterize the market equilibrium. In the  $z$  market, the equilibrium condition is a standard one, since all potential investors are

presumed to trade in this market.<sup>15</sup> The clearing condition is  $\sum_{i=1}^N q'_{iz} = 0$ , or equivalently,<sup>16</sup>  $\sum_{i=1}^N q_{iz} = \sum_{j=1}^M b_j Q_j$  (where, to recall,  $Q_j$  is the exogenously given supply of asset  $j$ ). A CAPM-like pricing condition is thus obtained by the use of (18).

Denote by  $N_j(p_j)$  the set of investors for whom condition (19) holds for asset  $j$  as a strict inequality, given the price  $p_j$ . Similarly denote by  $\bar{N}_j(p_j)$  the set of investors for whom (19) holds weakly (i.e.,  $\bar{N}_j(p_j)$  will include all the marginal investors in asset  $j$  when its price is  $p_j$ , in addition to  $N_j(p_j)$ ). The equilibrium price  $p_j$  can then be determined by the condition

$$(20) \quad \sum_{i \in N_j(p_j)} q_{ij} \leq Q_j \leq \sum_{i \in \bar{N}_j(p_j)} q_{ij}.$$

Indeed, if (20) holds for  $p_j$ , the left-hand inequality implies that the quantity demanded by the active intramarginal investor equals or falls short of the quantity supplied. If it falls short, then by the right-hand inequality, marginal investors will be willing to purchase the excess at the price  $p_j$ .<sup>17</sup>

The characterization of the market equilibrium in (20) is unsatisfactory in two respects. First, it involves inequality instead of the standard equality conditions; and second, it leaves the price only implicitly determined. To overcome these difficulties, two additional assumptions will be introduced.

In order to avoid inequalities, it is now assumed that investors can be considered as a continuum, and that personal characteristics vary continuously across investors. To enable us to obtain a more explicit specification, it is further assumed that investors are generally similar and that they differ only in one characteristic. Since differences in the

<sup>15</sup>The implicit assumption here is that the trade in  $z$  (in both directions) is costless to all the  $N$  potential investors, independently of whether they trade in any other risky asset or not.

<sup>16</sup>The assumption that individuals share the same assessments of the parameters  $b_j$  is necessary here. Otherwise, the equilibrium in the  $z$  market may depend on the composition of investors' portfolios.

<sup>17</sup>I ignore here the possibility that since the number of marginal investors is discrete, one of them may have to purchase less than his threshold quantity.

expectations  $E_i x_j$  seem to be the most pertinent here, it will be assumed that these are the only personal parameters in which investors differ from one another. Thus  $E_{iz} = E_z$ ,  $Var_i e_j = Var e_j$ ,  $Var_i z = Var z$ ,  $c_{ij} = c_j$ ,  $t_{ij} = t_j$ ,  $A_i = A$ , and, as a result,  $\pi_{iz} = \pi_z$ , and  $\phi_{ij} = \phi_j$ , for all  $i$ .<sup>18</sup>

Given the second simplification, it is quite obvious that the active purchasers of asset  $j$  will be the investors with the highest expectation  $E_i x_j$  about that asset. It is convenient therefore to denote by  $Ex_j(h)$  the expectation of that individual  $i$  whose subjective assessment  $E_i x_j$  ranks in the  $h$ th place among all investors. Ignoring the problem of ties, we thus obtain a ranking of the expectations of all potential investors,  $Ex_j(1) \geq Ex_j(2) \geq \dots \geq Ex_j(N)$ , and corresponding to it, a ranking of the respective individuals. By the first simplification, it is now assumed that  $Ex_j(h)$  can be considered as a continuous function in  $h$ .<sup>19</sup>

The problem of defining the set of active investors reduces now to that of defining their number  $n_j$ . If not all investors are active, condition (19) will hold at the margin with equality. Thus,

$$(21) \quad p_j(1+r+t_j) = Ex_j(n_j) - b_j\pi_z - \phi_j.$$

<sup>18</sup>It should be noted that we should generally expect the investor's relevant subjective estimate of an asset's variance to include some measure of his doubt concerning his own estimate of the mean. The more informed are thus likely to be characterized by their lower subjective variance estimates, rather than by their estimates of the mean. It is thus chiefly in order to simplify the exposition that I concentrate on differences in the assessments of the first moments of the probability distributions of assets. Alternative equilibrium conditions to those obtained below can be obtained for the case where investors differ in the parameters which are reflected in  $\phi_{ij}$ , when  $E_i x_j = E x_j$ ,  $t_{ij} = t_j$  for all  $i$ . If  $\phi_j(h)$  is constructed as increasing across investors (similar to the construction of  $Ex_j(h)$  below), then the equilibrium is given by  $p_j(1+r+t_j) - Ex_j + b_j\pi_z = \phi_j(n_j) = (Q_j/n_j) \bar{A}_j(n_j) \bar{Vare}_j(n_j)$ , where  $\bar{A}_j$  and  $\bar{Vare}_j$  are weighted harmonic averages of the  $n_j$  investors with the lowest  $\phi_{ij}$ .

<sup>19</sup>The assumption that  $Ex_j(h)$  is continuous is introduced here in order to guarantee the existence of a marginal investor  $n_j$  (for whom (19) holds with equality) in the case when not all investors are active. Discontinuities in  $Ex_j(n)$  may, however, occur; indeed the number  $N$  of potential investors may be considered as a point of discontinuity.

The clearing condition  $\sum_{i \in N_j} q_{ij} = Q_j$  can be written now by the aid of (18) as

$$(22) \quad p_j(1+r+t_j) = \bar{E}x_j(n_j) - b_j\pi_z - \frac{A}{n_j} Q_j Var e_j,$$

where  $\bar{E}x_j(h)$ , the average expectation function, is defined by  $\bar{E}x_j(h) = (1/h) \cdot \sum_{i=1}^h Ex_j(i)$ . The condition which determines the equilibrium premium  $\pi_z$ , given the assumption of similarity of investors, is

$$(23) \quad \pi_z = \frac{A}{N} \left( \sum_{j=1}^M b_j Q_j \right) Var z.$$

The characterization of the equilibrium in conditions (21) and (22) is a special case of (20).<sup>20</sup> Another possibility that has to be examined, though, is that *all* investors will choose to be active traders. In this case, condition (19) will hold with inequality for everyone and (21) will have to be replaced by

$$(24) \quad Ex_j(N) \geq p_j(1+r+t_j) + b_j\pi_z + \phi_j.$$

The clearing condition (22) remains intact, except that  $n_j$  is replaced by  $N$ . Using (23) as well, the clearing condition can then be rewritten as

$$(25) \quad p_j(1+r+t_j) = \bar{E}x_j(N) - \frac{A}{N} Cov \left( x_j, \sum_{k=1}^M Q_k x_k \right).$$

Equilibrium condition (25) is, in fact, the one generated by the CAPM (given the assumption that  $A_i = A$  and that expectations about variances and covariances are unanimous).

<sup>20</sup>If short sales are permitted, a condition supplementing (19) would be that  $j$  should be in  $S_i$  if  $\pi_{iz} - b_j\pi_{iz} < -\phi_{ij}$ . In the equilibrium conditions, we would have to distinguish the marginal purchaser  $n_j^l$  from the marginal short seller,  $N - n_j^s$ . An equation similar to (21) would have to be added to the effect that  $\pi_j^s(N - n_j^s) = b_j\pi_z - \phi_j^s$ ;  $n_j^l$  would have to be substituted in (21), and an average  $\bar{E}x_j(n_j^l, n_j^s)$  and some factor  $t_j(n_j^l, n_j^s)$  would replace  $\bar{E}x_j$  and  $t_j$  in (22).

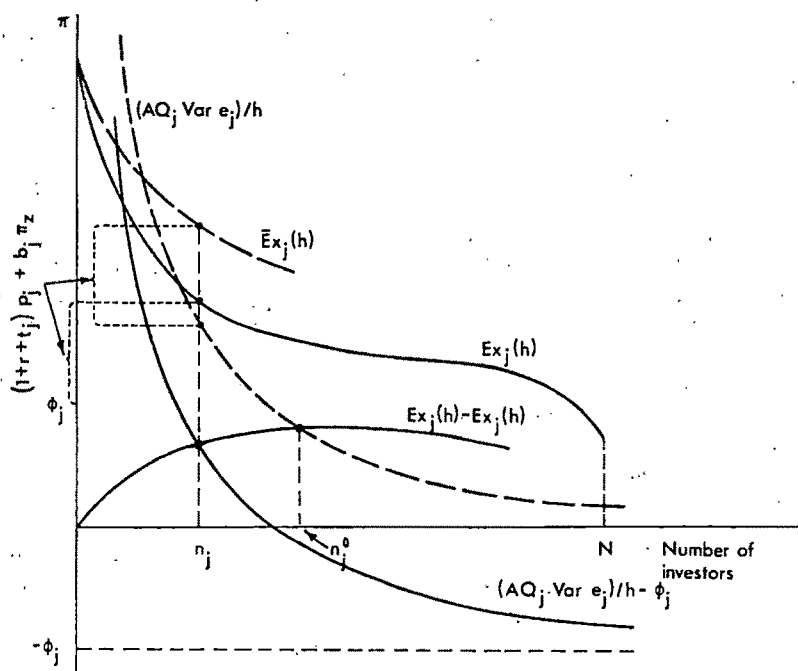


FIGURE 1. AN ASSET'S EXPECTATION SCHEDULE  
AND ITS EQUILIBRIUM PRICE AND NUMBER OF INVESTORS

If condition (24) is recognized as an integral part of the *CAPM* characterization (25), it becomes clear that the determination of asset prices may be closely interlinked with the identification of active investors. This link is much more obvious in the solution of (21)–(22) where some investors refrain from active purchase. While condition (21) adapts the equilibrium formula of my 1979 paper, condition (22) restates the *CAPM*-like equilibrium formula of Lintner (1969). As is quite obvious by now, neither of these two conditions should be considered as determining the equilibrium price— $p_j$  and  $n_j$  are to be considered as *jointly* determined.

The simultaneous conditions (21)–(22) illustrate the statement in the introduction that when opinion diverges the marginal-opinion and the average-opinion theories of asset prices may both be correct. The former theory, that proposed by Keynes, Williams, and Miller, may be interpreted as considering (21) to determine the equilibrium price. The latter theory, that of Lintner (1969) and Sharpe (1970), ascribes the same role to condition (22). Both theories, however, should

be regarded as complementary rather than conflicting.

A graphic representation of the joint solution for  $p_j$  and  $n_j$  in (21)–(22) is provided in Figure 1. The number of investors is depicted on the horizontal axis and a presumably typical expectation schedule  $Ex_j(h)$ , similar to the one drawn by Williams and Miller, is assumed. The equilibrium is obtained from the condition that at the margin,  $n_j$ , the difference  $\bar{Ex}_j(h) - Ex_j(h)$  between the expectations of the average and marginal investors should be equal to the difference between their respective specific-risk premia  $(AQ_j \text{Var } e_j)/h - \phi_j$ . Once  $n_j$  is determined, the equilibrium  $p_j$  can be obtained, as illustrated, by using either one of equations (21) and (22).

To obtain equations which explicitly determine each of the two endogenous variables, one has to specify the shape of the expectation schedule. A useful simple specification is the linear one,

$$(26) \quad Ex_j(h) = \hat{Ex}_j - \delta_j h.$$

In this case, if we define the parameter  $\eta_j$  by  $1 + \eta_j = (1 + Q_j \delta_j / c_j)^{1/2}$ , equations (21)–(22) can be solved to yield

$$(27) \quad n_j = \frac{\phi_j Q_j}{c_j (\eta_j + 2)} = \frac{\phi_j \eta_j}{\delta_j}$$

and

$$(28) \quad p_j = 1 / (1 + r + t_j) \times [\bar{E}x_j - b_j \pi_z - \phi_j (1 + \eta_j)].$$

The pricing formula (28), in which the fixed costs  $c_j$  and the degree of diversity of opinion (as measured by  $\delta_j$ ) affect the equilibrium price, clearly differs from the standard CAPM formula (25). It might seem that the existence of fixed purchase costs is the main source of the difference. However, this is not so. Figure 1 can be used to demonstrate that a reduction in the fixed purchase costs, which lowers  $\phi_j$  and thus shifts  $(AQ_j \text{Var } e_j) / h - \phi_j$  upward, will generally increase the number of active investors and simultaneously increase the equilibrium price. However, even in the absence of fixed purchase costs (when  $\phi_j = 0$  but short sales are still prohibitive), the number of active investors, identified in Figure 2 as  $n_j^0$ , may still fall short of the total number of potential investors,  $N$ . In this case, the equilibrium relations (21)–(22) will still apply. Indeed, if we consider the limit case  $c_j = 0$ , equations (27)–(28), for the case of a linear expectation schedule, may still imply a non-CAPM solution where

$$(1 + r + t_j) p_j = \bar{E}x_j - b_j \pi_z - (2AQ_j \delta_j \text{Var } e_j)^{1/2}$$

$$\text{and } n_j = (2AQ_j \text{Var } e_j / \delta_j)^{1/2}, \quad n_j \leq N.$$

The nonexistence of purchase costs (when short sales are restricted) is thus *not* a sufficient condition for the CAPM results to hold, since as long as there is sufficient divergence of opinion, it may still happen that not all

potential investors choose to be active.<sup>21</sup> Nonexistence of fixed purchase costs, combined with homogeneity of opinion (in the linear case,  $\delta_j = 0$ ) provide sufficient, though not necessary, conditions for the CAPM results to hold.<sup>22</sup> But do such conditions hold in reality, even as an approximation?

#### IV. Implications

##### A. Alternative Structures of Opinion

Various comparative-static exercises may be conducted in examining the effects of parameter changes on the market equilibrium. In the present context, it seems appropriate to concentrate on the structure of opinion by employing equilibrium conditions (21)–(22) and varying the shape of  $Ex_j(h)$  while keeping all else constant.

Consider first the change in opinion among nonactive investors illustrated in Figure 2, where (as in Figures 3 and 4) the original opinion schedule is denoted by  $E$ , the schedule  $\bar{E}$  denotes average opinion,  $\Delta$  stands for  $\phi_j + \bar{E}x_j(h) - Ex_j(h)$ , and  $H$  marks the hyperbola  $(Q_j A \text{Var } e) / h$ . The equilibrium  $n_j$  is obtained (in a minor variation of Figure 1) at the intersection of  $\Delta$  and  $H$ . For brevity,  $(1 + r + t_j)p_j + b_j \pi_z + \phi_j$  is denoted by  $\hat{p}_j$  and referred to as the price. The corresponding curves and parameters for the alternative expectation schedule are identified by primes.

In the alternative expectation schedule considered in Figure 2,  $E'x_j(h) = Ex_j(h)$  for

<sup>21</sup>Note that in the absence of homogeneity of opinion, the nonexistence of fixed transaction costs on purchases and short sales (implying  $\phi_j = \phi_j^s = 0$ ), may still not be sufficient to guarantee the CAPM formula. Existence of proportional transaction costs, even if only on short sales, may still result in a simultaneous-system solution for the price, the marginal purchaser and the marginal short seller, as described in the preceding note.

<sup>22</sup>Jarrow (1980) analyzed the direction of the bias between the prices which prevail in an imperfect market and the CAPM prices. He shows that when investors share the same view on assets' covariances (but not necessarily otherwise), the CAPM prices are higher than those which apply when short-sales restrictions are introduced. It should be recognized, however, that the introduction of fixed purchase costs as a form of market imperfection may reverse the direction of the difference since a rise in these costs (unlike a rise in fixed short-sales costs) results in lower equilibrium prices.

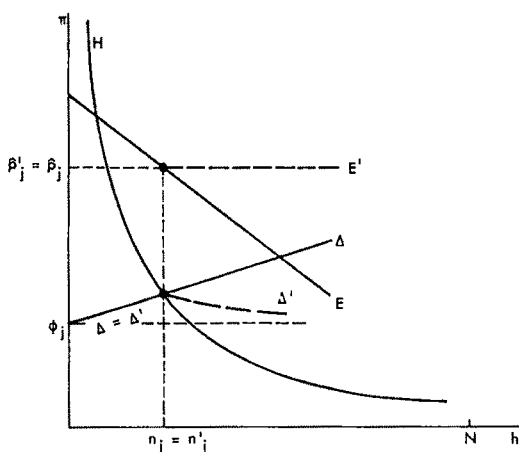


FIGURE 2. A CHANGE IN OUTSIDE OPINION

$h \leq n_j$  and  $E'x_j(h) = Ex_j(n_j)$  for  $h \geq n_j$ . The market equilibrium remains intact in spite of the fact that the expectations of some investors have changed. We may thus conclude—as did Williams—that a change of opinion among nontraders which is not large enough to induce them to become traders will have no effect at all.

Second, consider the case, depicted in Figure 3, where the original schedule is flat in the range of the marginal investor. In this case, a change in the expectations of intramarginal investors may affect only the number of active investors, leaving the price unchanged. Thus, provided there are enough marginal investors, a change in the opinion of active investors may also not be reflected in the price.

Third, consider the case (not illustrated) where a parallel shift in opinion takes place, that is,  $E'x_j(h) = Ex_j(h) + a$ . Because  $\bar{E}$  will also shift by the same amount,  $\Delta$  will remain intact, and the only change will be in the equilibrium price  $\hat{p}_j$  rising to  $\hat{p}_j + a$ . This parallel shift in opinion thus changes the price, but is not reflected in the pattern of trade at all, leaving the identity of the active investors and the volume of trade of each of them unaltered.

Next to be considered is a change in the degree of diversity of opinion. Here it is convenient to use the linear-expectations case (26) where the parameter  $\delta_j$  may be readily

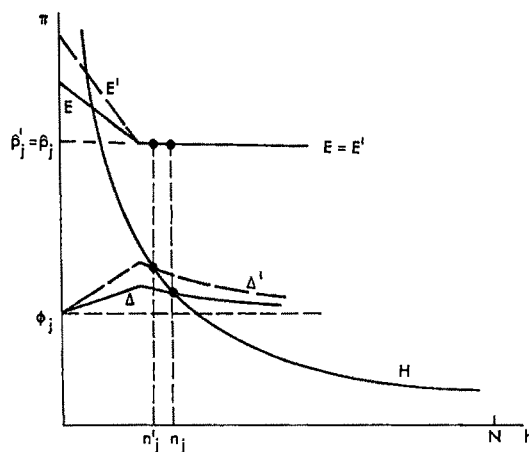


FIGURE 3. A CHANGE IN INTRAMARGINAL OPINION

identified as a measure of diversity. From (27) we may note that the equilibrium number of investors,  $n_j$ , is inversely related to  $\delta_j$ . A decrease in the diversity of opinion (perhaps resulting from wider dissemination of knowledge) thus results in an increase in the number of active investors.

The equilibrium price  $p_j$  depends on both the level of expectations (as measured for example by the intercept  $\bar{E}x_j$ ) and the diversity of opinion. Thus when rotating the expectations schedule, with changed  $\delta_j$ , it is important who is the pivot for the rotation (i.e., who is the investor  $h$  for whom  $E'x_j(h) = Ex_j(h)$ ). A rise in  $\delta_j$  which keeps  $\bar{E}x_j$  constant implies not only greater diversity of opinion, but also an overall lowering of expectations. On the other hand, a pivot beyond the marginal investor implies higher expectations for all active investors. A change of the latter type served as the basis for Miller's claim that greater divergence implies a higher equilibrium price (and a lower return).

It may be more appropriate to consider the pivot as somewhere in the middle range of active investors. Miller's proposition will still apply in the linear case if the pivot is to the right of the average investor  $n_j/2$ . Increased diversity will, however, have a contrary, negative effect on prices when it leaves the average opinion of active investors unchanged, so that  $\bar{E}'x_j(n_j) = \bar{E}x_j(n_j)$ .

The last result may be considered to provide a theoretical basis for the empirical findings of Irwin Friend, Randolph Westerfield, and Michael Granito, who used a survey of expectations about stock returns to find that higher average expected returns were associated with greater diversity. Assuming that the investors surveyed are representative of active investors, so that their average opinion  $\bar{E}x_j$  is representative of  $\bar{E}x_j(n_j)$ , equations (26) and (28) can be transformed to yield

$$(29) \quad \bar{E}x_j/p_j = (1 + r + t_j) \\ + (b_j/p_j)\pi_z + \left[ \frac{A}{2} \text{Var}(e_j/p_j) \right]^{1/2} \\ \times \left[ c_j^{1/2} + (c_j + Q_j\delta_j)^{1/2} \right].$$

In this equation the average expected return of active investors,  $\bar{E}x_j/p_j$ , is indeed positively related to the divergence of opinion about that average, as represented by  $\delta_j$ . The effects of fixed and proportional costs, and of the variance of specific return are also explicit in this formulation.

Finally, one can compare a decreasing expectations schedule with an alternative horizontal schedule where all investors share the same expectation  $E'x_j$ . This comparison, which can be considered as a special case of the preceding one, is singled out because of its importance, and is depicted in Figure 4.<sup>23</sup> The number of active investors is larger in the uniform case. However, the equilibrium price  $\hat{p}_j'$  depends on the level of  $E'x_j$ . The price will remain unchanged only if the uniform expectation  $E'x_j$  is equal to the expectation of the marginal investor  $Ex_j(n_j)$ . If  $E'x_j$  is equal to the expectation of any intramarginal investor (and in particular to that of the average active investor, as it is in Figure 4), the price would rise. Alterna-

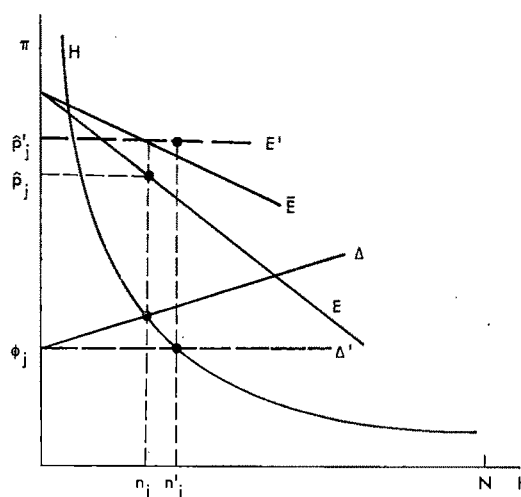


FIGURE 4. DIVERGING AND HOMOGENEOUS AVERAGE EXPECTATIONS

tively, the price would fall if  $E'x_j$  were equal to the average expectation of all investors  $Ex_j(N)$  and, at the same time,  $\bar{E}x_j(N) < \bar{E}x_j(n_j)$ .

These simple comparative-static exercises demonstrate the importance of the shape of the distribution of opinion among investors. They also demonstrate how misleading it may be to consider the average opinion of either actual or potential investors as "representative," that is, as representing homogeneous-equivalent expectations that, if held by everyone, would leave prices unchanged. While marginal opinion appears to play the role of homogeneous-equivalent expectations in these examples, this is clearly not the case in general, as can be recognized by examples with discontinuities at the margin (in particular at  $N$ ) where there may be no one to whom marginal conditions such as (21) apply, or by examples where investors differ in a number of characteristics (or when there are short sales) where many different investors may be marginal.

### B. On Informational Efficiency

The model discussed here referred to differences in investors' subjective expectations, but specifically avoided reference to any objective probabilities and to the infor-

<sup>23</sup>The case of uniform opinion can be solved by substituting  $\delta_j$  in (27) and (28). The equilibrium conditions reduce to those derived in my 1979 paper, where the CAPM-like condition (22) serves to identify the marginal investor while condition (21) is the one which determines the price.

mation which underlies expectations. In this framework it is straightforward to show that, given investors' expectations and given the structure of trading costs and the supply scale of assets, the allocation of risk achieved by the market is Pareto efficient.

Recently, however, an entirely different notion of efficiency, namely the informational efficiency of capital markets, has drawn much attention. This concept was defined by Eugene Fama as referring to a market in which prices "fully reflect all available information." Two alternative interpretations of this definition may be distinguished in the context of the two-period framework adopted here, *ex ante* and *ex post*. To check for informational efficiency on the former interpretation, one would examine what information is reflected in prices. On the *ex post* interpretation, second-period hindsight would be used to determine whether first-period prices of alternative assets were in some sense "correct" given (all) the then-available information.

The *ex post* interpretation of informational efficiency seems to be the one that has been more widely tested. The model used here sheds little light on it, however, since it does not consider the informational basis of expectations and therefore does not incorporate a notion of the correctness of expectations. It is thus beyond the scope of this paper to evaluate the claims for the *ex post* informational efficiency of the market or the contrasting claims of Williams and Miller that stocks "have a tendency to sell too high" (Williams, p. 29) or that asset prices "will exceed the willingness to pay of an investor with perfect information" (Miller, p. 1153). These claims are, indeed, based on the rather arbitrary assumption that "most people are right," so that  $\bar{E}x_j(N)$  is the "correct" expectation (and, in addition, on the assumption that  $\bar{E}x_j(N) < Ex_j(n_j)$ ).

The *ex ante* interpretation of informational efficiency avoids the need to specify the correctness of expectations. The ideal set here, however, is still quite ambiguous. A minimal requirement would be that *all* information gets reflected in the market price in some way. The *CAPM* equilibrium price (25)

meets this requirement since it integrates the opinions of all investors.

A more demanding requirement, suggested by Mark Rubinstein, is that prices would remain the same even if all individuals were to have identical expectations based on all the information available. In this case, the homogeneous-equivalent (or consensus) expectations would be the full-information expectation. To the extent that the average expectation  $\bar{E}x_j(N)$  of all investors is equal to the expectation that would be formed on the basis of all the information, we might interpret the *CAPM* price in (25) as meeting this requirement as well.<sup>24</sup>

An alternative requirement, suggested and analyzed by Sanford Grossman and Joseph Stiglitz, is that prices convey (are a sufficient statistic for) the entire information, in the sense that an outside investor could learn just as much about an asset from its equilibrium price as from all the available information. Grossman and Stiglitz demonstrate that if it is assumed that investors know the equilibrium price before submitting their final demands to the market auctioneer, then such informational efficiency would be impossible (unless all information is costlessly available to begin with).

The foregoing discussion illuminates, I believe, an important facet of *ex ante* informational efficiency, namely the effects of transaction costs and restrictions on short sales.<sup>25</sup>

<sup>24</sup>This statement implicitly assumes that investors would not revise their assessments of second moments if all information became available to them. Furthermore, the assumption that the market weighted average of investors' expectations is the proper (full information) aggregation of those expectations may not be valid. Independent information possessed by a few may be underweighted in comparison with the duplicated information possessed by many. When investors differ in their risk aversion, the market average would similarly underweigh the opinion of those with relatively higher risk aversion.

<sup>25</sup>An interesting early statement concerning information efficiency, which recognizes the importance of short sales, was provided by Alfred Marshall: "The private purchaser of railway shares may know nothing [about its prospects, the ability of its management and the propriety of its accounts], but he buys with the confidence that all such points have been scrutinized by many keen

Once it is recognized that such imperfections imply that not all investors will trade in an asset, it becomes evident that prices cannot reflect *all* information since there will exist individuals who have not revealed their information via trade—except by abstaining. Furthermore, even if it were assumed that the information of nonactive investors does not matter, once it is realized that prices and the identity of active investors are simultaneously determined, it is no longer clear what information is reflected or conveyed by equilibrium prices. For some of the examples examined here, it could indeed be argued that only the information of marginal investors is reflected in the price. More generally, however, the equilibrium price depends in a complex manner on the structure of information and trading costs across investors. Since full-information expectations should be independent of the distribution of private information among potential investors, and since homogeneous-equivalent expectations were shown to depend on the structure of opinion (and, thus, on the distribution of information), there can be no general equivalence between them.

### V. Conclusion

The model presented here incorporates imperfections in capital markets in the form of both trading costs and divergence of opinion among investors. It was shown that when (because of trading costs) not all investors trade in all assets, equilibrium prices and the identity of the investors trading in each asset are jointly determined. It was then demonstrated that, given divergence of opinion, equilibrium prices can be considered as determined simultaneously by the average and marginal investors. Thus the Keynes-Williams marginal-opinion theory and the Lintner-Sharpe average-opinion theory were shown to be complementary (rather than al-

ternative) characterizations of equilibrium asset prices.

An important implication of this result is that it is as a rule incorrect to regard a situation with heterogeneous opinion as representable by one in which all market participants share identical average expectations. The common "as if" assumption of the existence of a representative individual with homogeneous-equivalent expectations may thus provide an inadequate theory of capital markets, in much the same sense that the substitution of certainty equivalents for probability distributions provides an inadequate theory of the individual's behavior under uncertainty.

The framework suggested here, in which transaction costs exist and heterogeneity of investors is essential, may prove useful in the analysis of undeveloped, fragmented capital markets, and in examining the role of financial intermediaries in imperfect markets. I also believe that this framework can provide a basis for models in which the primary role of capital markets in allocating investment resources would be more realistically portrayed than in the currently available perfect market models.

### REFERENCES

- Arrow, Kenneth J., "The Role of Securities in the Optimal Allocation of Risk-Bearing," *Review of Economic Studies*, April 1964, 31, 91-96.
- Bohm-Bawerk, Eugen von, *Capital and Interest*, Vol. II, South Holland: Libertarian Press, 1959.
- Debreu, Gerard, *Theory of Value*, New Haven: Yale University Press, 1959.
- Diamond, Peter, "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *American Economic Review*, September 1967, 57, 759-76.
- Fama, Eugene F., "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, June 1978, 33, 902-17.
- Friend, Irwin, Westerfield, Randolph and Granito,

---

men with special knowledge, who are able and ready remorselessly to 'bear' the stock if they find in it any weak spot, which... had not been allowed for in making up its value" (p. 95; see, however, also his contrasting viewpoint on p. 9).

- Michael, "New Evidence on the Capital Asset Pricing Model," *Journal of Finance*, June 1978, 33, 903-17.
- Grossman, Sanford J. and Stiglitz, Joseph E., "Information and Competitive Price Systems," *American Economic Review Proceedings*, May 1976, 66, 246-53.
- Jarrow, Robert, "Heterogeneous Expectations, Restrictions on Short Sales, and Equilibrium Asset Prices," *Journal of Finance*, December 1980, 35, 1105-13.
- Keynes, John Maynard, "The General Theory of Employment," *Quarterly Journal of Economics*, February 1937, 51, 209-23.
- Lintner, John, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, February 1965, 47, 13-37.
- , "The Aggregation of Investors' Diverse Judgements and Preferences in Purely Competitive Markets," *Journal of Financial and Quantitative Analysis*, December 1969, 4, 347-400.
- Marshall, Alfred, *Money Credit and Commerce*, London: Macmillan, 1923.
- Mayshar, Joram, "Transaction Costs in a Model of Capital Market Equilibrium," *Journal of Political Economy*, August 1979, 87, 673-700.
- , "Transaction Costs and the Pricing of Assets," *Journal of Finance*, June 1981, 36, 583-97.
- Miller, Edward M., "Risk, Uncertainty, and Divergence of Opinion," *Journal of Finance*, September 1977, 32, 1151-68.
- Rubinstein, Mark, "Securities Market Efficiency in an Arrow-Debreu Economy," *American Economic Review*, December 1975, 65, 812-24.
- Sharpe, William F., "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, September 1964, 19, 425-42.
- , *Portfolio Theory and Capital Markets*, New York: McGraw Hill, 1970.
- Tobin, James, "Liquidity Preference as Behavior Towards Risk," *Review of Economic Studies*, February 1958, 26, 65-86.
- Williams, John Burr, *The Theory of Investment Value*, Amsterdam: North Holland, 1956 (reprint of 1938 ed.).

# An Essay on the Foundations of Friedman's Methodology

By WILLIAM J. FRAZER, JR. AND LAWRENCE A. BOLAND\*

Milton Friedman's famous essay on methodology (1953, pp. 3–43) has been presented as an instrumentalist argument for instrumentalism (Boland, 1979). Although Friedman has stated that such characterization of his methodology is "entirely correct," he also says (1979) that his views on methods and the philosophy of economics can be aligned with those of the philosopher Karl Popper. Many students of the philosophy of science would be possibly shocked by such a claim, since Popper has so often criticized and rejected both instrumentalism and Logical Positivism (1965). Thus, against such background, we present our examination of Friedman's claim. Evidence which we gathered to determine the extent of the relationship between Friedman's methodology and Popper's view of science will be presented. It will be argued that to a degree Friedman's claim can be supported.

The major point is that Friedman identifies with Popper for two reasons: 1) Both Friedman and Popper reject Positivism; 2) Friedman and his followers tend to argue indirectly in advancing a view by criticizing its alternatives. Presumably, we either accept Positivism (for example, Paul Samuelson, Carl Hempel, Rudolph Carnap, among others), or we do not. Obviously, this line of argument considers only two among a possibly larger number of options (a confrontation of competing views, as it were, as found in Friedman's view of hypothesis testing). A minor point is that Friedman does not see Popper dismissing instrumentalism completely. In this he is correct, since Popper only criticizes instrumentalism as a philosophy of science, while accepting it as a methodology for social policy.

Except for a brief statement that "Factual evidence can never 'prove' a hypothesis, it can only disprove it" (1953, p. 9), nothing in Friedman's essay *seems* to depend on Popper's philosophy of science. So, if there is an alignment of Friedman's view with Popper's, it will have to be found separately from the famous essay. Thus, we examine the prospect with reference to his academic and research backgrounds (including examples from his work), and to his association with Popper. A corollary purpose of our essay is to indicate how Friedman's view of the methodology of "positive economics" depends on his acceptance of some (but perhaps not all) of Popper's philosophical discussions of the nature and purpose of science.

As we indicate, Friedman's early orientation in methods had no immediate background in established economics. That the main and crucial elements of Friedman's work came from outside may be gleaned from a review of these elements and their antecedents. This emphasis reveals the following: 1) that Friedman has been incorrectly identified with Logical Positivism (Warren Gramm, pp. 169, 171, 175; Samuelson, 1963, pp. 82–83), a doctrine about the unity of science as formulated in Vienna in the 1920's; and 2) that Friedman is much less aligned than commonly presumed with the views of his Chicago mentors. In Friedman's economics, theories are partly arguments for alternative policies (and/or social reform). An interdependent system of analytic constructions (rather than isolated construction in economics) is used to generate hypotheses and for predicting the effects of alternative policies. Contrary to the conventional views of statistical induction and logical deductions from known true assumptions, as dealt with elsewhere (Boland, 1979, pp. 512–13; 1981), the purpose of economic theory for Friedman is prediction for purposes of testing and evaluating alternative policies. These policies over Friedman's career

\*The University of Florida and Simon Fraser University, respectively. We thank John Chant, Stephen Easton, Herbert Grubel, Zane Spindler, and James F. O'Conner for comments on an early draft. Milton Friedman was cooperative in answering questions we raised. We alone are responsible for the final product.

have been mainly those that followed from Keynesian and his own economics.

The rejection of standard philosophies of science is also encountered in Popper's work. For Popper, science is "trial and error" and "conjecture and refutation." There is no method which will guarantee or "prove" a successful trial or a correct conjecture. For Friedman, acceptance in the end is a matter of public debate and decision making.

# I. Logical Positivism vs. Friedman's Methodology

## A. The Essence of the 1953 Methodology Essay

The 1953 essay (Friedman, pp. 3-7) was to call attention to the great relevance of positive economics (empirical study) for normative economics ("what ought to be"). Given alternative policies A and B (for example, a Keynesian interest rate control policy vs. a Friedmanian monetary aggregates policy, Frazer, 1980, chs. 4; 6), the question was which policy should be selected. Such question of how we get from where we are to the policy goal was thus seen as an empirical question, as one of selecting the most useful theory among available competitors.<sup>1</sup>

In broad outline this theme of the relevance of positive for normative economics was foreshadowed in Friedman's dissertation with Simon Kuznets (1945) and explicitly dealt with in Friedman's thinly spread writings on methodology. Rose Friedman lists them: "Lange on Price Flexibility and Employment: A Methodological Criticism," (1946), "The Marshallian Demand Curve," (1949), "Methodology of Positive Economics," the lead essay in his *Essays in Positive Economics* (1953), and "Leon Walras and His Economic System" (1955), a review article of the English translation by William Jaffé of Walras' great treatise (1976, p. 20). To these we would add "Comment on 'A Test of an Econometric Model for the United States, 1921-1947,'" (1951) and *Monetary*

*Trends* (1982, ch. 3, and sections 2.3 and 6.2).<sup>2</sup>

The assertion about which controversy centered in Friedman's 1953 essay was that predictive success and not descriptive accuracy is the true test of a theory, "that there is no way of judging the realism of a theory except by the conformity of its predictions to observations" (Friedman, p. 20). The key passages from Friedman's essay start thus:

In so far as a theory can be said to have "assumptions" at all, and in so far as their "realism" can be judged independently of the validity of predictions, the relation between the significance of a theory and the "realism" of its "assumptions" is almost the opposite of that suggested by the view under criticism. Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions (in this sense).

[1953, pp. 14-15].

In a separate footnote on this last proposition, Friedman asserted: "The converse of the proposition does not of course hold: assumptions that are unrealistic (in this sense) do not guarantee a significant theory."

Friedman continues with reference to the proposition: "the more significant the theory, the more unrealistic the assumptions [in the sense of descriptive detail, a ledger of all the entries as it were]." He says:

The reason is simple. A hypothesis is important if it "explains" much by little, that is, if it abstracts the common and crucial elements from the mass of complex and detailed circumstances surrounding the phenomena to be explained and permits valid predictions on the basis of them alone. To be important, therefore, a hypothesis must

<sup>1</sup>This selection should not be confused with choosing the "best" theory; see Boland (1971).

<sup>2</sup>References to Friedman's and Anna Schwartz's 1982 work will be to sections of the book. The first number in the reference (section 2.5) is to the chapter and the second is to the section within that chapter.

be descriptively false in its assumptions; it takes account of, and accounts for, none of the many other attendant circumstances, since its very success shows them to be irrelevant for the phenomena to be explained.

To put this point less paradoxically, the relevant question to ask about the "assumptions" of a theory is not whether they are descriptively "realistic," for they never are, but whether they are sufficiently good approximations for the purpose at hand. And this question can be answered only by seeing whether the theory works, which means whether it yields sufficiently accurate predictions. The two supposedly independent tests thus reduce to one test. [p. 15]

What Friedman was against in these passages was the conventional view in the economics of the time, and in strong measure beyond it, that conclusions could be derived from assumptions and the truth or falsity of the derived statements would follow from the truth or falsity of the assumptions. Arguments about the detailed accuracy of the assumptions along with the logic was to establish in a sense, an empirically oriented theory.

### B. *The Defense of Friedman's Logic*

Contrary to many of the published critiques of Friedman's 1953 essay, it has been previously argued (Boland, 1979) that Friedman's essay is free of logical errors.<sup>3</sup> More-

<sup>3</sup>The methodology offered as positive economics by Friedman in this 1953 essay was to give rise to considerable controversy for over a decade initially and extending even beyond. The initial controversies focused on the role and realism of assumption (D. V. T. Bear and Daniel Orr, 1967; Gerald Garb, 1965; Tjalling Koopmans, 1947, 1957; Abba Lerner, 1965; Fritz Machlup, 1955, 1964; Gerald Massey, 1965; Jack Melitz, 1965; Ernest Nagel, 1963; Samuelson, 1963, 1964, 1965; Stanley Wong, 1973). Furthermore, the controversy over realism extends to the philosophical and methodological dimensions emerging from econometric structural equations methods in the line of descent from Jan Tinbergen (co-recipient of the first Nobel Prize in economics) through Nobel laureate Lawrence Klein (Frazer, 1973, pp. 287-311, 391-92; 1980, ch. 13 and App. to ch 13; Friedman, 1951, pp. 107-14).

over, Friedman's essay can be clearly understood by seeing it as an example of the view of science called "instrumentalism." Boland (1979) states that the essay says,

... that theories are convenient and useful ways of (logically) generating what have turned out to be true (or successful) predictions or conclusions. Instrumentalism is the primary methodological point of view expressed in Friedman's essay.

For those economists who see the object of science as finding the *one* true theory of the economy, their task cannot be simple. However, if the object of building or choosing theories (or models of theories) is only to have a theory or model that provides true predictions or conclusions, *a priori* truth of the assumption is not required *if* it is already known that the conclusions are true or acceptable ....

[pp. 508-09]

The key here is that Friedman is not trying to solve the philosophical problems that have plagued the natural sciences for many centuries. In his 1953 essay, Friedman is more concerned with the immediate practical problems of policymaking than with the philosophical problems that have troubled those who have been searching for centuries for *the* method of finding the *one* true or acceptable theory. Moreover, it should be noted that Friedman (as well as Popper) also rejects the search for *the* method. Generally, instrumentalists never find the study of methodology very "useful."

### C. *Logical Positivism and Popper's Critique*

Logical Positivism as formulated in Vienna gives special attention to unity in the methodology of physical science and economics, and to the language of science as typified by the work of Rudolph Carnap (Peter Achinstein and Stephen Baker; Popper, 1965, pp. 253-92). For Carnap, metaphysics is a sort of "pseudo-science." In this language of sciences every legitimate statement of science would be a well-formed formula, "while none

of the meta-physical theories would be expressible in it—either because the terminology was not available, or because there was no well-formed formula to express it” (Popper, 1965, p. 265).

From the view point of Logical Positivism, a true science can only be concerned with verifiable (i.e., positive) statements. But, since any positive theory can only be verified by reference to given conventions (i.e., basic statement) of the language of science, all verifications are contingent proofs. Any true science should thus avoid discussing truth in any absolute sense.<sup>4</sup> The unity of science is founded solely on the wide acceptance of the conventions. The most that we could ever hope for is the establishment of a theory’s logical consistency (for example, whether it is a well-formed formula). Popper calls this view of science “Conventionalism” to distinguish it from its predecessor “Inductivism” which allowed the possibility of proofs based only on positive evidence uncolored by conventions.

Popper criticizes Logical Positivism (i.e., Conventionalism) as follows: He identifies a problem which he thinks Logical Positivism is intended to solve—the Demarcation Problem. Namely, how do we distinguish science from metaphysics. Note, this is not Popper’s problem, but rather his view of what he thinks is behind Logical Positivism. Popper criticizes Logical Positivism on the grounds that verifiability is an inadequate solution to the Demarcation Problem since absolute verifiability is logically impossible.<sup>5</sup>

Popper notes an early view of the demarcation problem: “The most widely accepted view was that science was characterized by its *observational* basis, or by its *inductive method*, while pseudosciences and metaphysics were characterized by their *speculative method*...or by the fact that they operated with ‘*mental anticipations*’ something very similar to hypotheses” (1965, p. 225). Now this Popper was unable to accept in its

strict statement. He did not wish to draw the line too sharply nor as sharply as Carnap drew it. He proposed that the demarcation be “the *refutability* or *falsifiability* of a theoretical system.” The theoretical statements should expose themselves to criticisms of all kinds. Acceptance depended on the statements standing up to criticism and attempted “refutations” from an empirical point of view.

Moreover, some theories exposed themselves more than others and, in addition, “we remember that most of our scientific theories originate in myths” (Popper, 1965, pp. 256–57). Indeed there was to be room for “conjecture” (for speculating, theorizing, or predicting from incomplete data).

Superficially, Popper and Friedman both see an element of unity in the methodology of physical science and economics. Popper perversely sees not a positive foundation of conventions, but instead sees an essential role for (destructive) criticism in any ongoing scientific community. Friedman accepts this, but gives attention to the role of prediction on the part of one theory vis-à-vis another as a means of settling the dispute and achieving acceptable means of proceeding in the conduct of economic policy.<sup>6</sup> The need for predictions is not ruled out by Popper, since he argues that testability is an essential element of scientific theories and testability relies exclusively on the refutability of a theory’s predictions (1968, ch. 3). A refutation of a theory is accomplished by finding a prediction which failed and thus concluding, *modus tollens*, that the theory must be false (see Boland, 1979, p. 504).<sup>7</sup>

<sup>6</sup>This confrontation of bodies of theory is well illustrated by Friedman (1982; Frazer, 1982b). In ch. 2 and other parts of the Friedman-Schwartz book, we encounter the simple quantity theory of money, the Keynesian challenge to the quantity theory, and a Friedman-Schwartz theory of nominal income. There are the theoretical results to be juxtaposed in the testing of the respective theories to be sure, but the issues separating Friedman from the Keynesians come in a special way to be methodological as well as simply differences in theories.

<sup>7</sup>It should be noted that the emphasis on prediction, on refutation and the use of *modus tollens* requires logical consistency. For this reason many have been led to consider Popper’s view as a variant of Logical Positiv-

<sup>4</sup>This conception of science is inconsistent, of course. It would have us not talk of truth, but, at the same time, it is claimed to be a true conception of science!

<sup>5</sup>This is the primary purpose of his famous book (1968).

The common error of identifying Friedman with Logical Positivism arises for a possible variety of reasons: because the term Positivism suggests an identity with "positive" economics, because of some identification of the monetarist Karl Brunner with Carnap (Brunner, 1967; Frazer, 1973, pp. 388-89), because Friedman held the view about the similarity of methods in the physical sciences and economics, and/or because of Carnap's presence in the United States at the University of Chicago.

On the false relation drawn between Positivism and Friedman, Friedman himself notes (in taped discussion, 1979) that Rudolf Carnap taught at Chicago for many years, in the Philosophy Department. "You see," Friedman says, "he was a refugee from Vienna, so many were after the Hitler era, and he was at Chicago during the 1940s and 50s at the same time Hayek [the Nobel Laureate in economics] was there." Carnap and Hayek, it seems, had been acquainted since the early 1920's, although they apparently were not close friends in either Vienna or in Chicago, and Hayek himself was critical of the Vienna Circle since student days. In any case, Friedman expresses doubt that Logical Positivism had any influence on his thinking, and he further notes that "Hayek and I have always differed on methodology, very fundamentally." On F. A. Hayek's influence on himself, Friedman says: "I think I have been much influenced by Hayek in respect of political economy, if you want, with respect to policy, but in this area, he and I have had arguments over many years about this, and we've never been able to agree. I think that if you look at his writings on methodology, you'll find that they have a wholly different flavor than my writings on methodology" (1979).<sup>8</sup>

## II. Friedman's Approach to Economics

### A. *The Background of Friedman's Economics*

Following an undergraduate major in mathematics and offers of scholarships in mathematics and economics, Friedman's background consisted of the following: 1) a school year at the University of Chicago with attention devoted to course work in statistics under Henry Schultz, and in political economy under Frank Knight, Henry Simons, and Jacob Viner; 2) a school year at Columbia with course work under Harold Hotelling in both mathematical statistics and mathematical economics, with James Angell in monetary economics, and with Wesley Mitchell in the history of thought and business cycles; 3) and continued empirical work at Wesley Mitchell's National Bureau of Economic Research, 1937-40.

These backgrounds were indeed highly diverse and unique. Mathematical statistics at the time was taught only at three institutions in the United States (Kenneth Arrow), and empirical research at the National Bureau was on an unparalleled scale with carefully defined antecedents going back to the 1890's at Chicago with J. Laurence Laughlin, Thorstein Veblen, and John Dewey. In the period until 1946, when Friedman arrived at Chicago as an associate professor, he was highly involved in mathematical statistics through Hotelling's influence, and mainly in the teaching of either statistics and/or business cycles. There was a school year as assistant on a research project for Henry Schultz, two years in Washington D.C. that resulted in a monumental piece of pure statistics (Friedman, 1937; Myles Hollander and Douglas Wolfe, 1973), and thirty-one months with Hotelling's war-related Statistical Research

ism—particularly so because one can recognize that "finding a prediction which failed" implies the acceptance of conventions concerning rules of evidence (Joseph Agassi, 1966, pp. 23-27; Mark Blaug, 1980, p. 19).

<sup>8</sup>On the difference in the methodology of the Austrian School, on the one hand, and Friedman, on the other, see Arthur Kemp (1978, pp. 16-17). As early as the fall of 1940, Friedman was to note disagreement with Hayek

on methodology. At that time Friedman had received a letter from N. Keyfitz, Dominion Bureau of Statistics, Ottawa, Canada. In the letter Keyfitz was referring to Friedman's review of Jan Tinbergen's *Business Cycles in the United States*, and he makes an aside remark about Hayek. Friedman replies thus (1940): "I have never been able to understand his dogmatic insistence on the proposition that statistical data could never verify economic hypotheses."

Group (SRG) at Columbia University (W. Allen Wallis). Continuing from SRG days, there was an association with Leonard Savage, the "subjective" or "personalistic" probabilist (Savage; Wallis), and with Karl Popper, the philosopher (1965; 1966; 1968; 1972).

As the foregoing may suggest, Friedman had considerable background for an economist of his day and beyond, in diverse areas in which no other single economist ventured in comparable degree.<sup>9</sup> However, there was no formal exposure to any unified philosophical doctrine of science in his early years except under Frank Knight at the University of Chicago. Knight's was an antiempirical view of economics. He held instead a complex philosophy of economics as an assumption oriented science (see Warren Samuels, pp. 59, 61–62, 346–47), but Friedman was to depart dramatically, by 180 degrees, as it were.

From Hotelling, Friedman received an introduction to the analytical significance of time rates of change as early as 1933–34, and from Hotelling, as brought forward in National Bureau work and economic theory, Friedman was to receive an orientation toward transitory and permanent (or quasi-permanent) time paths (Friedman and Kuznets, 1945, pp. 331–32, fn. 13; Horace Secrist).<sup>10</sup> Of the war-related Statistical Re-

search Group on the Columbia University campus. Friedman said in taped discussion, "There was probably the best statistical research group—there was the greatest collection of statisticians from this country that was ever assembled in one place, in terms of sheer intellectual ability, sheer ability in statistics, mathematical statistics." As a part of this Columbia experience and later, Friedman was to get a philosophy of statistics from Savage and to hold after that what may be called a Bayesian view rather than a classical one, though he never hesitated to use classical methods in analyses of data exclusive of their philosophical underpinnings.

Now the foregoing diverse influences combine in Friedman's hands to move the economics of static, isolated constructs and conventional method in a new direction. In that new direction are matters of time rates of change rather than levels, and of economists operating within a complex framework of theories (an engine of analysis, as it were, say as distinct from a limited view of isolated constructions in economics).<sup>11</sup> The use of the mechanics of economics as an engine of

---

economics was never to be the same again. Another example of the early controversies would be the one where Friedman and David Meiselman sought to test Keynesian theory by juxtaposing money and investment multipliers (Frazer, 1973, pp. 98–111). Still another would center about demand for money studies (Frazer, 1967, especially pp. 139–85; 170).

<sup>9</sup>Allen Wallis came closer than any other economist of the period to sharing Friedman's background as to depth and particular diversity of academic experiences. Wallis's year at Chicago in 1934–35 and the next year at Columbia with Hotelling follow with a two-year lag Friedman's own student year at the respective universities. Further, Wallis had a brief assignment at the National Bureau of Economic Research before moving to Stanford University and then to wartime activities. A large difference in the two individuals, however, was Wallis's strong attraction and suitability for administrative work over research and writing. The opportunities for further development after the wartime experiences surely made a difference in their respective careers. The crucial one in Friedman's development was his becoming heir to the monetary portion of Wesley Mitchell's grand design for the National Bureau.

<sup>10</sup>Friedman's use of rates of change and NBER turning points generated enormous controversy as had the discussion of axioms. This time rate of change controversy is reviewed elsewhere (Frazer, 1973, pp. 46–53). Controversy, it seems, always accompanied Friedman's early uses of the simple method and the subject of

<sup>11</sup>Some who are wedded to the treatment of levels in textbooks rather than to time rates of change as a part of the mechanics (theory) of economics will argue that time rates of change are just a tool. Following this line of argument, and recognizing that the basic Keynesian and price theory constructions all have their own mathematical underpinnings (Frazer 1980, chs. 5, 7, 8, in the first case, and chs. 15, 16, 17, in the second), all we are left with are mathematical tools. What Friedman does with time rates of change in stock and flow quantities, vis-à-vis levels, is to redirect the way the basic static constructions are cast with reference to the prospect of generating fruitful empirical hypotheses vis-à-vis the scant likelihood of the empirical estimation of static elasticities at points on static schedules. Examples come in Part C below. Not only does the introduction of time rates of change bring the historically static constructions more in accord with observed reality, but it does so without rejecting a role for the constructions. In addition, the movement toward time rates of change and other aspects of Friedman's approach helps in dealing with the statistical problems of heteroscedasticity, multicollinearity, and episodic changes.

analysis was to generate fruitful empirical notions.<sup>12</sup>

One may argue, as we do, that Friedman was trying to express in his 1953 essay (pp. 3–43) essential elements of a scientific methodology that admitted the foregoing influences and set the stage for monumental empirical research that had been planned at the National Bureau over thirty years ago (Friedman, 1949; 1979). In any case, at the organizational meeting of the Mont Pelerin Society in 1947, in Switzerland, Friedman met Karl Popper. From that meeting and later ones, Friedman developed an interest in Popper and his work. Popper's influence as expressed in *The Open Society and its Enemies* (1966) and *Conjectures and Refutations* (1965) are present in the formation of Friedman's thinking.

#### B. *The Indirect Method: A Reliance on a Complex Framework of Unprovable Theories*

The major thrust of Popper's view of science is the denial of any *direct* method of acquiring or proving (for example, by induction) the true theories of economics or any of the natural sciences. Most philosophers will agree with this, but many see Popper's judgments as a denial of *true* theories. Popper argues, however, that basing science on the denial of *true* theories is a defeatist position—the essence of the “conventionalist twist” (1965, p. 37). Conventionalism eschews empirical truth and instead focuses on the logical truth of complex (Walrasian) models of simultaneous solution of equilibria. Friedman does not choose to pursue such philosophical defeatism; yet he accepts Popper's

judgment on the nonexistence of a direct method. One can argue, as Friedman does, that the conventionalist methodology requires too much detail for practical purposes. This approach to detail is offered as if each little relevant bit of detail can be separated out as to its effect via *ceteris paribus* treatments, made a part of policy argument with the prospect of government control, and then offered to us as independent parts to be manipulated by government.

Supposedly, by using the body of standard economic analysis rather than the complex Walrasian approach of adding detail, simpler and more empirically refutable statements about the real world can be obtained. Examples of Friedman's approach follow.

#### C. *Examples of the Indirect Method*

Examples of Friedman's indirect methods occur almost without distinction between a priori constructions (such as Marshall's demand curve and Keynes's liquidity preference schedule) and statistical methods. In chronological order, examples may include Friedman on the Marshallian demand curve (1953, pp. 47–99), the econometric problem of identification (see Frazer, 1980, Appendix 1), Philip Cagan's approach to hyperinflation (1956, pp. 25–117), Friedman's consumption function (1957), and a stable relationship plus liquidity preference with overshooting (1968, pp. 11–27).

Though not limited to liquidity preference, hypothesis testing in the case of liquidity preference explicitly takes on the form of testing simple quantity theory, Keynesian and Friedmanian views of money demand against one another. There are, in other words, competing hypotheses. The testing may even take on a more general form of public debate and possible acceptance (or rejection) in the national policy sphere (Frazer, 1982b).

*The Demand Curve.* Knight had rejected the prospect of economics being an empirical science as a part of his reactions to a literal interpretation of so-called “economic man” concepts such as the measurement of utility. Friedman was consequently to avoid directly confronting these concepts, including his early essay on the Marshallian demand curve (1953, pp. 47–99). In that essay he drew

<sup>12</sup>The term “economic theory” in reference to courses and subject matter in economics often connotes what we refer to here as the mechanics of economics. The reference is to the common crosses (Marshallian, Keynesian, and so on), functions (consumption, utility, production, and so on), and the underlying mathematics of the main bodies of price monetary and theory (Frazer, 1980). In Friedman's hands, the bodies of analysis are being used to generate indirectly empirical statements that may be fruitful. As indicated below, this approach will differ from that of simply addressing directly slope parameters of schedules and hoping to be able to estimate the parameters and identify the schedules empirically, as was the case with Keynesians and David Laidler (Frazer, 1980, ch. 13; and Laidler, 1977).

the distinction between "the ordinary demand curve" and the "compensated demand curve," by addressing the *ceteris paribus* assumption. In the ordinary case, utility is locked up in the assumption about "all other things," but, in the second case, compensation for the loss of the utility of income is allowed for as the price of the commodity in question varies. There was no prospect for direct measurement of utility in connection with Marshall's curve, but Friedman was able to proceed indirectly to relate prices in Marshall's markets to a price index and then to indexation, that is, to measuring losses (or gains) to a household resulting from changes in the price index (the inverse of the purchasing power of income). (See Frazer, 1980, Appendixes 1; 16, and section 6.4.)

*Identification.* Friedman did not attempt any direct approach to the econometrician's notion of the identification of supply and demand schedules for money balances as found in Laidler (1977, pp. 114–17) and Ronald Teigen (1965).<sup>13</sup> Instead, on the supply side of the market, the money stock is a controlled variable fixed for the "economic actors" as a whole and in various ways (by the gold standard, monetary authorities, institutions, financial sophistication of the society). On the demand side, the money stock becomes a given variable to which economic actors may adjust readily by altering the rate of spending and hence the velocity of money and prices until nominal income is again in balance with the part of income the actors wish to hold in money balances. Monetary disturbances impacting on the supply side, however, can and do change

independently of changes in money demand. If this generalization were not so (i.e., "if (a) the quantity of money supplied were a function of the same variables as the quantity demanded and (b) the supply function was as stable over time and place as the demand function"), then "observed data on the quantity of money, nominal and real, and on the variables affecting the quantities of money supplied and demanded, would simply record random perturbations about the intersection of the stable demand and supply functions" (Friedman and Schwartz, 1982, section 2.3).

*Cagan's Essay.* Cagan's essay (1956) entitled "The Monetary Dynamics of Hyperinflation" was done under Friedman's direction as a dissertation. "This statistical combination with theory," Cagan said in his taped discussion (1978), "is definitely Friedman's.... I see it that way, because there was no one else in Chicago with that tradition, with that emphasis." As a statement of Friedman's style of working, Cagan also said:

You develop a theory and you develop the implications and then you hypothesize. In that essay [Friedman, 1953, pp. 3–43] I think he was trying to express the style of research, but he formalized it, tried to develop it philosophically, and I guess got into a lot of trouble, because then the philosophers started nitpicking about that and the other things....

As to the question of whether the methodology of *Positive Economics* was still the way to proceed, Cagan answered thus:

I still do, I always thought that was one of the most important things I learned from Friedman. It's not an easy thing to do, because you have to take a theory and work it around in such a way that it gives some implications. You don't set it up in such a way that you are clearly going to learn something because you are not. Not in the sense of whether the equation fits or not, because you can make almost any equation fit almost anything with enough variables in it, the way economic data go up and down.... [What

<sup>13</sup>In the initial essay on positive economics, Friedman is explicitly dealing with the economist's ability to list factors influencing demand and those affecting supply, such that the two sets of factors are independent of one another. He notes, however, that "the generalization is not always valid. For example, it is not valid for the day-to-day fluctuations of prices in a primarily speculative market" (1953, p. 8). Even supply and demand schedules for labor come to have some interdependence (Frazer 1980, section 17.4; Friedman, 1976, ch. 12). In the latter cases, the concepts of supply and demand "can still be used and may not be entirely pointless; they are still 'right' but clearly less useful ... because they have no meaningful empirical counterpart" (Friedman, 1953, p. 8).

one finds instead of this fitting of equations is] a special way of developing things... There is an idea there that's in the back of my mind that I felt he was emphasizing.

*The Consumption Function.* Going back to Hotelling (Friedman-Kuznets, 1945, pp. 331–32; Secrist, 1934) and the National Bureau, a central distinction Friedman uses in his approach is that between measured and permanent income, where the difference is transitory income (another name for the business cycle as a fluctuation about trend). Though not necessarily identical to measured and permanent, Friedman also draws a parallel between actual and anticipated magnitudes. Permanent income is thus anticipated income over a secular-trend time path. Friedman's main accomplishment in his consumption function work (1957) was to introduce permanent income into the function (most simply,  $C = kY_p$ , where the factor of proportionality  $k$  may shift with some other variables).

The mathematics of the permanent income expression are present in Friedman's approach (1957, pp. 143–45). Analytical results can be related to aggregate data through a special form of the adaptive expectations model. With respect to early efforts at the fitting process, explicit mention is made to Cagan's study above (Friedman, 1957, p. 145).

We end up via the foregoing routes with the following: the capacity to impose the permanent income consumption function on an ordinary Keynesian income/expenditure plane in a meaningful way, and with a reconciliation of otherwise apparent incongruous statistical results. On the Keynesian plane with income, expenditure, and consumption, on the vertical axis, and income to the factors of production as well as permanent income, on the horizontal axis, we may impose the new function along with a simple consumption relation,  $C = a + bY$ ,  $0 < dC/dY < 1$ . Through the manipulation of the adaptive expectations model, one can isolate the simple function as approximated by  $C = a + bY$ . This simple function now, however, is shifting with time (Frazer, 1980, section 9.4).

From the point of view of explanatory power, Friedman's approach reconciles earlier known results from budget (cross-section) studies with the finding that the ratio of saving to income ( $S/Y$ ) was constant over long periods of time. This would be exclusive of course of later governmental policies with respect to income redistribution that may have altered the ratio by favoring consumption more than saving out of income.

*The Stable Relation plus Liquidity Preference.* Friedman sought stable empirical relationships as a part of economic science with an emphasis on prediction. He finds this in one case as a slight variation of a relation stated in his initial restatement of the quantity theory of money (1956, p. 11). As restated and studied empirically in *Monetary Trends* (1982, section 2.4, equation (7)), there are real money balances on one side of the equation, and there are other variables, on the other. These include real income, a liquidity measure, expected rates of return for four different classes of assets, and a "portmanteau" variable.

Friedman's interest in this foregoing, stable relationship, however, is *not* in the effect of real magnitudes on real magnitudes. Rather it is in the interaction between real and monetary magnitudes. The main a priori construction for considering these is the liquidity preference building block as found in the economics of J. M. Keynes and Keynesian economics, except that it is recast along dynamic lines with time trends, transitory changes, and "overshooting" (Friedman and Schwartz, 1982, section 2.7).

In its static form, the liquidity preference construction has "the" interest rate (or its excess over the yield on money balances) on its vertical axis, and the quantity of money on the horizontal axis. There is an inverse relation (the liquidity preference schedule) defined by an equilateral hyperbola. The position of the asymptote to the hyperbola is dependent on nominal income, and a vertical money supply line is imposed on the graph. The construction gets recast, as stated, but it can also be related to Keynesian economics and Friedman's challenge to the Keynesians. Present in the construction, as described, is a ratio of income to the money stock. This

velocity ratio plays a key role in the old quantity theory of money, in Keynesian economics via the presumption of independence between the numerator and the denominator during the time at which the economics applies, and in Friedman's analysis of the demand for money. In Friedman's hands, as in those of Keynes, there is a portfolio of assets where, at equilibrium, rates of return on additions to the various classes of assets are equal (Frazer, 1980, section 3.3). These rates of return appear in the stable relationship but as expected, trend-path rates (Friedman and Schwartz, 1982, section 2.4, equation (7)).

In Keynesian theory, as reviewed by Friedman and Schwartz (1982), income is independent of the money stock, except that changes in the rate of interest may work to influence capital spending. Via varying the money stock (and sometimes going directly to open market operations), the Keynesians envision an inverse variation in the money stock and the rate of interest. In the Keynesians's hands, the demand for money is unstable at high rates of interest and infinitely elastic at low rates of interest. Hence control of the money stock is an ineffective means of controlling the interest rate in Keynesian economics.

In Friedman's theory, the time frames concerning transitory and permanent magnitudes are brought into his use of the liquidity preference construction. In addition, there is a NBER chronology and a shift of emphasis from levels in the stock and flow quantities to time rates of change. Equilibrium as a permanent magnitude of one where the velocity ratio ( $Y/M$ ) is constant, with denominator, numerator, and real output moving at the same rate. This comes up with the use of logarithms and the discussion of adjustment processes in response to a shift in monetary growth from one fixed rate to another, and later with reference to a host of effects including combined liquidity and loanable funds effects, an intermediate income effect, and a long-run inflationary expectations effect (Friedman-Schwartz, 1982, section 10.1).

We need not labor these. Where expectations operate with a short lag, the expectations effect takes place almost immediately

and the traditional short-run effects mostly get submerged.<sup>14</sup> A main point is that accelerated growth in the money stock disturbs the balance of money "desired" by economic actors in relation to that determined by the central bank. Spending and therefore income ( $Y$ ) in the liquidity preference model accelerate with "overshooting" and interest rates rise rather than decline. The Friedmanian time frame is different, and the linkages and adjustment process are different from the counterparts in Keynesian economics.

From the point of view of a priori method, the Keynesian approach above is quite conventional. There is a short run in which the demand schedule holds in place, and in which an interest elasticity in the static sense may be estimated. Indeed, there is the implicit notion that the static schedule can be estimated empirically, or at least serve as a guide to actions in the real world. If some relation—such as one for the quantity of money to interest rates and income—is stable in the empirical sense, then fine. If not, then the analytical problem is one of adding variables to represent possibly omitted influences. In the end, the models can become quite large (Frazer, 1973, ch. 14).

In contrast to the above, Friedman proceeds indirectly, both from the points of view of a priori and statistical methods. Friedman and Schwartz say in reference to velocity and the demand for money (1982, section 6.2), "Our ultimate objective is to explain the behavior of velocity, which is to say, of the

<sup>14</sup>We encounter at this point in Friedman's work an empirical finding of special importance, namely: statistical results suggest "that the period of experience on which expectations were based shortened drastically after the mid-1960s" (p. 478). A so-called initial impact effect of accelerated money growth on interest rates (a negative correlation) began to be overwhelmed more readily by the inflationary expectations effect (positive correlation) even for periods as short as a quarter (Friedman-Schwartz, 1982, section 10.9; and fn. 107). They had expected the inflation-rate effect to be the major one for their cycle phases, but their original theoretical speculation was that the liquidity effect would be important within cycles and "largely to average out for ... cycle phases." What this finding suggests in effect is a role in the very short time frame for a purer psychological analysis, something most economists have avoided in the past (Frazer, 1980, section 21.3). (See also Frazer, 1982a.)

quantity of money demanded, that takes account simultaneously of all the variables affecting velocity" (p. 215). They say their objective is best served by proceeding "indirectly." This means examining key variables one or two at a time with reference to the hypotheses generated by the theory (say, their use of the liquidity preference construction). "We believe," they say, "the indirect approach yields insights that cannot be obtained from the more sweeping approach ['the prevailing fashion in econometric work'].... that multiple correlations with many variables are almost impossible to interpret correctly unless they are backed by more intensive investigations of smaller sets of variables" (p. 215). This is in contrast to proceeding immediately to compute multiple regressions "including all variables that can reasonably be regarded as relevant."

#### *D. Friedman's Version of Popper's Approach to Science*

About the time that Friedman met Popper, Popper was preparing an address, "Predictions and Prophecy in the Social Sciences" (see Popper, 1965, ch. 16). This address is probably the only single work by Popper that might be seen to give direct support to Friedman's methodology. In the address, Popper distinguished between "unconditional" scientific predictions and "conditional" scientific predictions. The former are possible in physical sciences because the system in question is a stationary and repetitive system, as in the old historical case of the solar system and the prediction of eclipses. The solar system, indeed, is viewable as being "isolated from the influences of other mechanical systems by immense regions of empty space and therefore relatively free of interference from outside" (Popper 1965, pp. 339-40). Commenting on these repetitive systems, Popper says, they are "special cases where scientific prediction becomes particularly impressive—but that is all" (p. 340).

In social history, on the other hand, "evidence is far more difficult to interpret." Along these lines, Friedman says:

It is frequently complex and always indirect and incomplete. Its collection

is often arduous, and its interpretation generally requires subtle analysis and involved chains of reasoning, which seldom carry real conviction. The denial to economics of the dramatic and direct evidence of the "crucial" experiment does hinder the adequate testing of hypotheses; but this is much less significant than the difficulty it places in the way of achieving a reasonably prompt and wide consensus on the conclusions justified by the available evidence. It renders the weeding-out of unsuccessful hypotheses slow and difficult. They are seldom downed for good and are always cropping up again. [1953, pp. 10-11]

Now Friedman says, as does Popper, that this added difficulty of the social science area does not make the natural sciences and economics fundamentally different with respect to method. Friedman's position on this similarity between the social and physical sciences is summarized by Rose Friedman:

His position is that the social sciences do not differ ... from the physical sciences and that those who contend otherwise largely do so because they misconstrue the nature of the physical sciences. No science can generate meaningful hypotheses that are "true" in any other sense than that they are provisionally accepted because no hypothesis generating more accurate (or less costly) predictions has been discovered. This position is squarely at odds with the praxiological methodological view of Ludwig von Mises and his disciples—even though they and my husband are very close together in their ideological approach to economic organization. [1976, p. 20]

All this, of course, emphasizes the importance of evaluating hypotheses. Where Popper is concerned with a philosophy of science involving universal or timeless evaluations, Friedman's approach is more immediate, grounded in the need to solve immediate practical problems of policymaking.

### E. *The Dependence of Predictions on Conditional Hypotheses*

The difficulty with reasoning from assumptions and then turning back to establish the detailed and descriptive accuracy of the assumptions, Friedman says, is that we presume better evidence is available but that the test of the conformity of assumptions to reality "is a test of the validity of the hypothesis *different* from or *additional* to the test by implications" (1953, p. 14). This has caused "much mischief," he says, and it has promoted misunderstanding.

As viewed by both Friedman and Popper (1965, pp. 339–40), it would appear that prediction is a hallmark of the physical sciences, but their reasons are quite different. For Popper, predictions are a source of potential refutations. Popper admits that as a matter of logic "all theoretical sciences are predicting sciences," and that some social sciences are theoretical. He does not equate the latter with "historical prophecies," as found in Marxist doctrine and in John Stuart Mill, however, but rather makes the distinction between "unconditional predictions" and "scientific predictions." The conditional predictions may be equivalent to unconditional ones "only if they apply to systems which can be described as well isolated, stationary and recurrent." The solar system is such a system where conditional predictions become unconditional ones.<sup>15</sup>

In the initial essay on methodology, Friedman means by prediction mainly prediction from conditional hypotheses. Such comes up

initially and on ordinary interpretation as a prediction about quantity demanded as price declines, other things being equal, as with a Marshallian demand schedule.<sup>16</sup> In major later works, however, expectations in the markets play a more prominent role and methods that obtain statistical results in a very indirect way become more prominent as examples in Part C above should indicate. The predictions become more "unqualified" and depend on the use of conditional predictions to make more unqualified predictions. Also, attention to the unintended consequences of our actions and the body of economics, lead to the formulation of "practical technological rules" that state *what we cannot do*. An example as found in Popper would be that "'You cannot, without increasing productivity, raise the real income of the working population' and 'you cannot equalize equalize real income and at the same time raise productivity'" (1965, p. 343).<sup>17</sup>

Examples of unqualified predictions in Friedman's work would include the following: the ratio of consumption to income is a permanent or quasi-permanent magnitude (i.e., may be treated as a constant or variable constant, in the long run); there is no money illusion (or "irrationality" with respect to inflation, to use a newer term) in the long run (Frazer, 1980, section 17.4; Friedman, 1976, ch. 12); proceeding from equilibrium (permanent or quasi-permanent) time paths for money growth of a given percent per year, nominal income of the same rate, and an interest rate as would be given, then raising the money growth rate by two percentage points will increase the inflation rate by two percentage points and the interest

<sup>15</sup> On unconditional and conditional predictions, Popper may also be quoted:

Unconditional scientific predictions can sometimes be derived from these conditional scientific predictions, together with historical statements which assert that the conditions in question are fulfilled. [From these premises we can obtain the unconditional prediction by the *modus ponens*.] If a physician has diagnosed scarlet fever then he may, with the help of the conditional predictions of his science, make the unconditional prediction that his patient will develop a rash of a certain kind. But it is possible, of course, to make such unconditional prophecies without any such justification in a theoretical science, or—in other words—in scientific conditional predictions. They may be based, for example, on a dream—and by some accident they may even come true. [1965, p. 339]

<sup>16</sup> Confer fn 13.

<sup>17</sup> A restatement of this rule in dynamic terms would go as follows: With real income per worker growing along a trend line at the rate  $X$ , you cannot raise the real income per worker above the trend line without faster growth in productivity, and you cannot equalize the distribution of real income (say, smaller saving-to-income ratio) and at the same time maintain or raise productivity above its trend line. Further, Friedman (1953, p. 23) notes that "negative statements can generally be made with greater confidence than positive statements" in discussion of methodology with reference to empirical science.

rate by the same amount (Frazer, 1980, section 5.4; Friedman, 1968; Friedman-Schwartz, 1982).

Also found in Friedman and in Popper is the notion that the main thrust of economics (or science generally) is toward solutions to problems. Popper in particular treats science as an ongoing program of bringing "rational criticism to bear on the problems that face us, and on the solutions advanced by the various parties" (1965, p. 337). He speaks of being armed "with the weapons of a *critic of methods*." The essence of Popper's view (1965, p. vii) is that science is an ongoing process whereby knowledge advances through our learning by our mistakes and misguided attempts to solve problems.

Friedman is also concerned with the problems of a society. Addressing them on occasion compounds the controversy over methodology, though Friedman was by no means to abandon logic in doing so. The testing of one theory against another may take the ultimate form of social debate, as in the case of Keynesians vis-à-vis monetarists in Britain's Thatcher government (Frazer, 1982b). The hypothesis/public debate may be kept distinct from the ideological debate except for the need for open discussions and debate. However, in the methodological sphere, confidence in the accepted hypothesis resides in the methodology rather than in "trusted authority," and policy discussion may get extended to such issues as rules to limit authority versus discretionary policy (Frazer, 1973, pp. 349-81).<sup>18</sup> In brief, the Friedman version of a Popperian view of science and the development of economics depends on science being problem oriented. Both science and economics are alleged to thrive on free discussion and critical debate.

<sup>18</sup>Examples of this compounding of controversy—1) over methodology, 2) over the acceptable hypothesis, 3) over ideology, and 4) over rules vs. authority (Frazer, 1973, pp. 349-81)—are found in extended confrontations between Franco Modigliani and Friedman, between Paul Samuelson and Friedman, and between others and Friedman on other occasions. The two opponents noted explicitly were in opposing positions on all four counts with respect to the compounding of controversy, yet boundaries for the issues in dispute were never clearly set in their confrontations.

### III. Indirect Philosophy: The Limits of Popper's Support

The only support that Popper's philosophy provides Friedman is indirect. Direct support of Friedman's instrumentalism is certainly denied by Popper's extensive critiques of instrumentalism in general (1965, chs. 3; 6). The indirect support is provided by Popper's rejection of both the problem of induction and the philosophy of Logical Positivism (1968; 1972). The question then is how can Friedman dismiss Popper's rejection of instrumentalism? We think this can be easily explained on two grounds. First, philosophically, Friedman's views on methodology are instrumentalist in the shorter run where policymakers reside (including in the philosophical context less than a quarter in duration, cycles, trends, and long swings). The span of time is in contradistinction to an infinitely distant future where some ultimately true theory may reside, as found in the imagination of some philosophers. Second, Popper's approach to the shorter-run, practical problems of social policy on close examination turns out to be one variant of instrumentalism, one which he calls "piecemeal engineering."

A closer examination of Popper's critique of instrumentalism will show that there is room for an instrumentalist approach to economics if one is confined to practical problems. Popper summarizes his criticism of instrumentalism as follows:

...Instrumentalism can be formulated as the thesis that scientific theories—the theories of the so-called "pure" sciences—are nothing but computation rules (or inference rules); of the same character, fundamentally, as the computation rules of the so-called "applied" sciences. (One might even formulate it as the thesis that "pure" science is a misnomer, and that all science is "applied".)

Now my reply to instrumentalism consists in showing that there are profound differences between "pure" theories and technological computation rules, and that instrumentalism can give a perfect description of these rules

but is quite unable to account for the difference between them and the theories.... [1965, p. 111]

So long as one is only concerned with immediate "applied" science and eschews "pure" theories, Popper's criticism of instrumentalism loses its force. Thus Popper leaves room for Friedman to dismiss the criticism of instrumentalism: Popper says, "... For instrumental purposes of practical application a theory may continue to be used *even after its refutation*, within the limits of its applicability: an astronomer who believes that Newton's theory has turned out to be false will not hesitate to apply its formalism within the limits of its applicability..." (1965, p. 113).

To the extent that Friedman is concerned with only practical applications, he can thus even find support in Popper's recognition that instrumentalism may be an appropriate methodology for some problems. Instrumentalism as found in Friedman's approach emphasizes a relevance of empirical data (i.e., positive evidence) in deciding what we ought to do (i.e., normative procedure). Deciding what we ought to do is considered a practical problem of choosing between alternative theories. Thus, for those who choose to extend the instrumentalism of Friedman's essay on methodology, explanation is not entirely cast aside for mere prediction.

## REFERENCES

- Achinstein, Peter and Baker, Stephen E., *The Legacy of Logical Positivism: Studies in the Philosophy of Science*, Baltimore: Johns Hopkins University Press, 1969.
- Agassi, Joseph, "Sensationalism," *Mind*, January 1966, 75, 1-24.
- Arrow, Kenneth J., "Harold Hotelling at Columbia," correspondence, November 21, 1979.
- Bear, D. V. T. and Orr, Daniel, "Logic and Expediency in Economic Theorizing," *Journal of Political Economy*, April 1967, 75, 188-96.
- Blaug, Mark, *The Methodology of Economics*, New York: Cambridge University Press, 1980.
- Boland, Lawrence A., "Methodology as an Exercise in Economic Analysis," *Philosophy of Science*, March 1971, 38, 105-17.
- \_\_\_\_\_, "A Critique of Friedman's Critics," *Journal of Economic Literature*, June 1979, 17, 503-22.
- \_\_\_\_\_, "On the Futility of Criticizing the Neoclassical Maximization Hypothesis," *American Economic Review*, December 1981, 71, 1031-36.
- Brunner, Karl, "The Controversy Between 'Quantity-Theory' and 'Keynesian-Theory': A Case Study on the Importance of Appropriate Rules for the Competitive Market in Ideas and Beliefs," *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, June 1967, 103, 173-90.
- Cagan, Philip, "The Monetary Dynamics of Hyperinflation," in Milton Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1956, 25-117.
- \_\_\_\_\_, "The Methodology of Positive Economics," taped discussion, November 9, 1978.
- Frazer, William, *The Demand For Money*, Cleveland: World Publishing Company, 1967.
- \_\_\_\_\_, *Crisis in Economic Theory*, Gainesville: University of Florida Press, 1973.
- \_\_\_\_\_, *Expectations, Forecasting, and Control: A Provisional Textbook of Macroeconomics*, Vols. I; II, Lanham: University Press of America, 1980.
- \_\_\_\_\_, (1982a) "The Velocity-Interest Rate Association: Inflation, Accelerated Inflation, and Uncertainty," *Economic Notes*, No. 1, 1982, 11, 144-53.
- \_\_\_\_\_, (1982b) "Milton Friedman and Thatcher's Monetarist Experience," *Journal of Economic Issues*, June 1982, 16, 525-33.
- Friedman, Milton, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, December 1937, 32, 675-701.
- \_\_\_\_\_, "Letter to N. Keyfitz," October 3, 1940.
- \_\_\_\_\_, "Money and Banking," in Arthur F. Burns, *Wesley Mitchell and the National*

- Bureau, Twenty-Ninth Annual Report of the National Bureau of Economic Research, 1949, 80-81.
- \_\_\_\_\_, "Wesley Clair Mitchell as an Economist Theorist," *Journal of Political Economy*, December 1950, 58, 463-95.
- \_\_\_\_\_, "Comment on 'A Test of an Econometric Model for the United States, 1921-1947,'" in *Conference on Business Cycles*, New York: National Bureau of Economic Research, 1951.
- \_\_\_\_\_, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.
- \_\_\_\_\_, "Leon Walras and His Economic System," *American Economic Review*, December 1955, 45, 900-09.
- \_\_\_\_\_, "The Quantity Theory of Money—A Restatement," in *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1956.
- \_\_\_\_\_, *A Theory of the Consumption Function*, Princeton: Princeton University Press, 1957.
- \_\_\_\_\_, "Factors Affecting the Level of Interest Rates," in Donald P. Jacobs and Richard T. Pratt, eds., *Savings and Residential Financing: 1968 Conference Proceedings*, Chicago: U.S. Savings and Loan League, 1968, 11-27.
- \_\_\_\_\_, *Price Theory*, Chicago: Aldine Publishing Company, 1976.
- \_\_\_\_\_, "Taped Discussion with Milton Friedman," Stanford University, February 6, 1979.
- \_\_\_\_\_, and Kuznets, Simon, *Income from Independent Professional Practice*, New York: National Bureau of Economic Research, 1945.
- \_\_\_\_\_, and Schwartz, Anna Jacobson, *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices and Interest Rates, 1867-1975*, Chicago: University of Chicago Press, 1982.
- Friedman, Rose D., "Milton Friedman: Husband and Colleague—(2) The Beginning of a Career," *Oriental Economist*, June 1976, 44, 18-22.
- Garb, Gerald, "Professor Samuelson on Theory and Realism: Comment," *American Economic Review*, December 1965, 55, 1151-53.
- Gramm, Warren S., "Chicago Economics: From Individualism True to Individualism False," in Warren J. Samuels, ed., *The Chicago School of Political Economy*, East Lansing: Association for Evolutionary Economics, 1976.
- Hollander, Myles and Wolfe, Douglas A., *Non-parametric Statistical Methods*, New York: John Wiley & Sons, 1973.
- Hotelling, Harold, "The Economics of Exhaustible Resources," *Journal of Political Economy*, April 1931, 39, 137-75.
- Kemp, Arthur, "The Political Economy of Milton Friedman," *Modern Age*, Winter 1978, 22, 8-17.
- Koopmans, Tjalling C., "Measurement with Theory," *Review of Economics and Statistics*, August 1947, 29, 161-72.
- \_\_\_\_\_, *Three Essays on the State of Economic Science*, New York: McGraw-Hill, 1957.
- Laidler, David, *The Demand for Money: Theories and Evidence*, 2d ed., New York: Harper & Row, 1977.
- Lerner, Abba P., "Professor Samuelson on Theory and Realism: Comment," *American Economic Review*, December 1965, 55, 1153-55.
- Machlup, Fritz, "The Problem of Verification in Economics," *Southern Economic Journal*, July 1955, 22, 1-21.
- \_\_\_\_\_, "Professor Samuelson on Theory and Realism," *American Economic Review*, September 1964, 54, 733-35.
- Massey, Gerald J., "Professor Samuelson on Theory and Realism: Comment," *American Economic Review*, December 1965, 55, 1155-63.
- Melitz, Jack, "Friedman and Machlup on the Significance of Testing Economic Assumptions," *Journal of Political Economy*, February 1965, 73, 37-60.
- Nagel, Ernest, "Assumptions in Economic Theory," *American Economic Review Proceedings*, May 1963, 53, 211-19.
- Popper, Karl R., *Conjectures and Refutations; The Growth of Scientific Knowledge*, 2d ed., New York: Harper & Row, 1965.
- \_\_\_\_\_, *The Open Society and Its Enemies*, Vols. I, II; 1st eds., rev., Princeton: Princeton, New Jersey Press, 1966.
- \_\_\_\_\_, *The Logic of Scientific Discovery*, 2d

- ed., New York: Harper & Row, 1968.
- , *Objective Knowledge*, Oxford: Oxford University Press, 1972.
- Samuels, Warren J., *The Chicago School of Political Economy*, East Lansing: Association for Evolutionary Economics, 1976.
- Samuelson, Paul A., "Problems of Methodology: Discussion," *American Economic Review Proceedings*, May 1963, 53, 231–36.
- , "Theory and Realism: A Reply," *American Economic Review*, September 1964, 44, 736–39.
- , "Professor Samuelson on Theory and Realism: Reply," *American Economic Review*, December 1965, 55, 1164–72.
- Savage, Leonard J., *The Foundation of Statistics*, New York: Dover Publications 1972.
- Secrist, Horace, "Two Letters to the Editor," *Journal of the American Statistical Association: Papers and Proceedings*, 1934, 29, 196–98; 200.
- Teigen, Ronald L., "The Demand for and Supply of Money," in *Readings in Money, National Income, and Stabilization Policy*, Homewood: Richard D. Irwin, 1965.
- Wallis, W. Allen, "The Statistical Research Group, 1942–1945," *Journal of the American Statistical Association*, June 1980, 75, 320–30.
- Wilber, Charles K. and Wisman, John D., "The Chicago School: Positivism or Ideal Type," in Warren J. Samuels, ed., *The Chicago School of Political Economy*, East Lansing: Association for Evolutionary Economics, 1976.
- Wong, Stanley, "The *F*-Twist and the Methodology of Paul Samuelson," *American Economic Review*, June 1973, 63, 312–25.

# Enlistments in the All-Volunteer Force: A Military Personnel Supply Model and Its Forecasts

By COLIN ASH, BERNARD UDIS, AND ROBERT F. MCNOWN\*

One test of the validity of a scientific hypothesis is its forecasting accuracy. Thus one test of the hypothesis that the market can be used to allocate manpower to defense, as to any other occupation, is its ability to predict voluntary enlistments. In our earlier paper, a simple model of accessions and enlistments to the U.S. armed forces was estimated, the conclusion being that the all-volunteer force (AVF) "is an experiment in market economics which, far from having failed, has not yet been put to the test" (McNown et al., 1980, p. 130).

Although one may not be particularly sanguine about the prospects for testing the market experiment itself, one can at least expose the hypothesis to further scrutiny. This paper reports the results of an accuracy analysis of forecasts generated by the model, and, since the results on the whole are encouraging and the topic remains of considerable interest to the public and to policy-makers alike,<sup>1</sup> also presents updated estimates of the personnel supply elasticities.

\*Lecturer in economics, University of Reading, England; professor and associate professor of economics, University of Colorado-Boulder, respectively. Ash worked on this paper while on leave at the University of Colorado. During 1982-1983, Udis is visiting professor of economics at the USAF Academy. We gratefully acknowledge research support from the University of Colorado. We are also indebted to Tally O'Donnell for excellent research and computing assistance. Colleagues Malcolm Dowling, Philip Graves, Jane Lillydahl, and Nicholas Schrock made valuable comments on an earlier draft of this paper; they are not responsible for any remaining errors or omissions.

<sup>1</sup>On January 23, 1980, in his State of the Union Address, President Carter asked Congress to reinstitute Selective Service registration as a part of America's response to the Soviet invasion of Afghanistan. President Carter's request was approved by Congress on June 25. Nineteen-year-old men began registration on Monday, July 21; twenty-year-old men began a week later.

## I. A Military Personnel Supply Model

The model borrows heavily from the work of Richard Cooper (1977), Anthony Fisher (1969), and Glenn Withers (1979). The underlying theory may be summarized briefly as follows. Each individual, traditionally a male over the minimum school-leaving age, faces a choice between civilian and military occupations, and there exists a military reservation wage at which each individual is indifferent between the two. This reservation wage is equal to the individual's highest alternative discounted earnings stream in civilian life, plus or minus a compensating differential reflecting the individual's relative taste for the nonpecuniary aspects of military service. Individuals may be arrayed according to their military reservation wage, creating a frequency distribution which is the joint distribution of civilian earnings and relative tastes. For a given level of military pay, all individuals with a lower reservation wage will enlist. Enlistment supply is thus defined by the cumulative frequency distribution of reservation military wages to be a general function of military and civilian earnings and tastes, *ceteris paribus*, as follows:

$$(1) \quad A^*/P = f(WM, \mu_{WC}, \sigma_{WC}, \mu_T, \sigma_T),$$

where  $A^*$  denotes applications to enlist, and  $P$ ,  $WM$ ,  $\mu_{WC}$ ,  $\sigma_{WC}$ ,  $\mu_T$ , and  $\sigma_T$  denote, respectively, the relevant population, the military pay of enlisted men (ranks E-1 to E-4 including basic pay, quarters, subsistence allowance, and tax advantages), the mean and standard deviation of expected civilian earnings, and the mean and standard deviation of relative tastes. The specific properties of the supply function will depend on the form of the joint frequency distribution of earnings

and tastes, on which there is no direct evidence.

Two further simplifying assumptions are made: first, that there is no change over time in either the variance of civilian earnings or of tastes; second, that the form of individual preference functions and their distribution across the population supports a relative pay hypothesis, and hence the inclusion of the ratio of civilian to military pay in the supply function.<sup>2</sup>

Previous empirical studies of enlistment in the *AVF* have included the civilian unemployment rate as an additional explanatory variable. Until relatively recently the theoretical justification for doing so has not been obvious. Here, however, the inclusion of the civilian unemployment rate is prompted by recent contributions to the theory of market disequilibrium, which emphasize (usually at the macroeconomic level) that "trade at false prices" in one market may impose quantity constraints on agents inducing or compelling them to revise their trading plans elsewhere.<sup>3</sup> If the real wage in the civilian labor market does not instantaneously adjust downwards in response to excess supply, actual unemployment is likely to result, forcing job hunters into the recruitment office. It is assumed in the present study that individuals' perceptions of likely employment constraints in the civilian labor market are determined by the corresponding unemployment rate.

It is also possible that the decision to enlist voluntarily is influenced by the likelihood of being drafted. In the absence of demand constraints, the enlistee can choose which service branch to enter; the draftee cannot. And, the experience of the latter years of the Vietnam War notwithstanding, the existence of the draft may signify a national call to arms, swelling the number of true volunteers. Hence a draft threat variable is included: although induction data do exist

by service branch, it is assumed that the relevant draft threat leading to draft-induced enlistments is the probability of induction into *any* service branch.

The function which we estimate is, therefore,

$$(2) \quad A_i^j/P_i = f(\bar{W}C_i/WM^j, U_i, \pi, T).$$

The variables are defined as follows:

$A_i^j$  = the number of nonprior service male accessions from the  $i$ th racial group into the  $j$ th branch of the armed forces.

$P_i$  = the population of 18–19-year-old males in the  $i$ th racial group, the primary source of nonprior accessions.

$WC_i$  = median civilian pay for 18–19-year-old males in the  $i$ th racial group.

$WM^j$  = average military pay of enlisted men, ranks E-1 to E-4, including basic pay, quarters, subsistence allowance, and tax advantages.

$U_i$  = the unemployment rate of 18–19-year-old males in the  $i$ th racial group.

$\pi$  = the probability of not being inducted into the armed forces, measured by one minus the total number of Department of Defense (DOD) inductions divided by  $P_i$ .

$T$  = a time trend intended to proxy a systematic change in tastes with respect to military service.

Complete descriptions of variables and data sources are given in our earlier paper (pp. 114–16), as are the results of experiments with alternative functional forms. The relative pay variable ( $WC_i/WM^j$ ) is entered into the equation with a one-period lag. This avoids a potential source of simultaneity and is consistent with lags observed by others.<sup>4</sup>

Although the theoretical model specifies the dependent variable to be applications to enlist, it will be noted that our empirical work is formulated around the accession rate and the actual enlistment rate. Data on applications to enlist are not readily available,

<sup>2</sup>In future work we intend to test the alternative, absolute pay hypothesis; at present we are unaware of any empirical evidence which would lead us to support it in preference to relative pay. See, for example, Withers.

<sup>3</sup>This is not the place to attempt a survey of a growth area in the literature, particularly when at least two are already available; see E. Roy Weintraub (1979) and Eleanor Moses (1980).

<sup>4</sup>See in particular David Grissmer (1979).

nor are disaggregated data on enlistments by race into each of the four services, Army, Navy, Marines, and Air Force. Accessions are the sum of voluntary enlistments and inductions (or draftees), and our study overlaps the years of the Vietnam draft. It is important to emphasize that the use of accessions rather than enlistments as our measure of supply in no way weakens the analysis. Given the above definition of accessions, it can readily be shown that a linear enlistment equation can be expressed as a simple linear transformation of the corresponding equation for accessions (see our earlier paper, pp. 120–25).<sup>5</sup> Although none of the key coefficients are altered by this transformation, the associated elasticities, when computed at the sample means, will differ according to the mean values of the accession and enlistment rate variables. Elasticities computed at the terminal values of the variables, when accessions equalled enlistments, are identical.<sup>6</sup> The use of total DOD inductions in constructing the draft threat variable in Army and Marine Corps regressions invalidates the precise algebraic relation between enlistment and accession equations discussed above. Nonetheless the estimated coefficients for the pay and unemployment variables do not differ significantly between the two dependent variable specifications.

The model was estimated for each armed service branch, both white and nonwhite racial groups, and for the DOD totals over the sample period 1967II–1976II using semianual data. Estimation by two-stage least squares prevented problems of potential simultaneity stemming from the unemployment component of both the dependent variable and the unemployment rate.

By way of summary, our evidence points to rather lower pay elasticities than had been previously estimated, no significant effect of unemployment on recruitment, a positive stimulus to voluntary enlistments from the

draft, and a weak but pervasive change in tastes away from military service.

The evidence on the lack of an unemployment effect on accessions is overwhelming. In none of twenty regressions is the unemployment variable significant at even the 35 percent significance level.<sup>7</sup> The standard errors of the unemployment coefficients are greater in absolute value than the estimated coefficients, in many cases more than three times as great. In seven of the twenty cases, the estimated coefficient has the incorrect (negative) sign. In view of the overall lack of explanatory power for this variable, these incorrect signs should be taken as further evidence of insignificance rather than an indication of model misspecification.

These results are consistent with previous analyses of the unemployment effect on enlistments. Cooper found a significant unemployment coefficient in his DOD enlistment equation (p. 168), but the implied elasticity is quite low; that is, .2 when evaluated at the mean. Fisher's estimated coefficient is statistically insignificant (p. 248). Alan Fechter's alternative models yielded highly unstable coefficient estimates, many with wrong signs and none with *t*-values greater than 2.0 (1979, pp. 93–95). Wither's unemployment coefficients for the United States are statistically significant but wrong-signed (pp. 125–28). David Grissmer's analysis suggests an explanation for these results. He also finds extremely low elasticities (.09 to .18) based upon marginally significant coefficients for mental category I–III volunteers. However, when this group is disaggregated, mental category I–II high-school graduates show moderate elasticities (.44 to .48) based on highly significant coefficients. Grissmer argues (pp. 106–08) that the armed forces increase their selectivity during periods of high unemployment, cutting back on non-high-school graduates while maintaining desired recruitment levels through higher rates of unemployment-induced enlistments from the more select group. This could certainly account for the weak unemployment effect observed for total volunteers.

<sup>5</sup>Experiments with alternative functional forms as well as evidence from other studies led us to select a linear supply function.

<sup>6</sup>For the Navy and Air Force, which accepted no inductees, the enlistment and accession elasticities are always identical.

<sup>7</sup>See our earlier paper, Table 2, pp. 121–23.

While our results on the effect of unemployment on accession are consistent with those presented by others, our conclusions on the pay elasticities vary somewhat from previous studies. Most work has indicated a pay elasticity greater than 1.0, and the Gates Commission employed an estimate of 1.25 in their projections of accessions following the end of the draft.<sup>8</sup> Our estimate of the pay elasticity of total DOD male accessions (or enlistments), evaluated at 1976 values of the variables, is .86. For the individual branches, only the regressions for the Navy show pay elasticities greater than 1.0 (1.015 for all males, when evaluated at the sample means), but none of the estimated elasticities for any racial group or any military branch are as large as the Gates Commission estimate. There is little variation between services in the enlistment pay elasticity. In terms of the end-period elasticities, the Navy is the highest and the Army the lowest. Such variation as exists reflects differences in volunteers' preferences toward the services, including different perception of the return over and above pay, differences in recruiting effort and its effectiveness, and different degrees of demand constraint. (See our earlier paper, Table 2, pp. 121-23.)

In comparing our pay elasticities with those of other studies, most of the differences can be accounted for by differences in data and model specification. The study closest to this in both data and model specification is the study by Fisher, who finds a no-draft elasticity estimate of .74 (p. 249). Both Fisher and Fechter use a draft-threat variable, as in this study, rather than attempting to define the dependent variable in terms of true volunteers. Fechter, in the static, relative pay version of his model, estimates elasticities which range from .64 to .88 (p. 94). Departures from this specification, with inclusion of a lagged dependent variable or absolute rather than relative pay variables, raise the pay elasticity estimate, in some cases, above unity (see p. 99). His "best" model, a static

version with pay in absolute terms, yields a pay elasticity of 1.12. In a comparative study of international manpower supply, Withers finds relative pay elasticities lower than our own, even allowing for the fact that his reported results refer to initial impact rather than long-run elasticities: 0.28 for the U.S. Army and 0.52 for all defense services (pp. 124-26).

The number of "true volunteers" is employed by both Grissmer and Cooper as the dependent variable. The number of true volunteers during the years of the draft lottery is given by the number of enlistees too young to have lottery numbers plus  $365/125$  times the number of enlistees with lottery numbers greater than 240. (See Cooper, p. 194.) This construction simply assumes a proportional number of true volunteers among those with lottery numbers under 240 as with those with numbers above 240. While such a procedure does control for draft-induced enlistments, it does not account for the negative effect the draft could have on enlistments. In addition to inducing enlistments through the threat of induction, the draft may have the opposite effect of discouraging enlistments because of the hostility toward the military generated by the existence of the draft itself. In other words, the probability of induction may still affect the rate of enlistments even among true volunteers. The elimination of the draft at the end of 1972 may have changed the attitude of candidates for enlistment and thus shifted the personnel supply function. If this is the case, then some of the growth in enlistments by true volunteers, which Cooper and Grissmer have attributed to pay increases, may be the result of the gradual reduction and eventual termination of military inductions. The Cooper and Grissmer studies which employ true volunteers consequently estimate pay elasticities generally higher than those which we have found. Cooper's pay elasticity estimates range between .95 and 1.23 (pp. 167-68), and Grissmer's range from .9 to 1.35 (p. 108).

Since the pay elasticity of personnel supply is an important policy issue, it is unfortunate that estimates of its value are sensitive to econometric issues which are difficult to re-

<sup>8</sup>See the President's Commission on an All-Volunteer Force, *The Report...*, p. 207. This Commission was chaired by Thomas F. Gates, and the *Report* is popularly known as the Gates' Commission Report.

solve theoretically or statistically. Estimated elasticities on the low end of the range of reported values come from studies, such as our own, which employ relative rather than absolute pay variables and define the dependent variable in terms of all enlistments rather than true volunteers. The resulting elasticities are generally less than one, indicating a given percentage change in the desired rate of enlistments requires a greater percentage increase in military pay.

## II. Retrospective Forecasts of Voluntary Enlistments

For the most part, the usual summary statistics pertaining to our estimated supply functions are satisfactory. Thus if the criterion for success is a satisfactory explanation of past enlistments, we claim modest success for our economic model. A much stiffer test however is to require a good track record for forecasts beyond the sample period over which the model was estimated. The model is therefore used to generate two sets of forecasts which are then exposed to a formal accuracy analysis.

The first set of forecasts are made using the accessions equations reported in our earlier paper. Given these parameter estimates, a multiperiod chain of forecasts can be generated for given values of the explanatory variables; that is, the predictive performance of the model is evaluated in isolation, immune from possible additional errors in forecasting the paths of, for example, relative military pay and civilian unemployment. The period covered by our accuracy analysis begins in the first half of 1977 and ends with the second half of 1979, six pairs of forecast and outcome data in all. (Throughout these years there was no draft, so the draft-threat  $\pi$  is constant at unity.)

We also generate a second set of forecasts for the same six half-years. In principle, ignoring important practical problems of lags in obtaining data and errors in provisional estimates, it would be possible for genuine forecasters to reestimate the equations every six months, as new outcome data became available. Applying these sequentially updated equations to data on the explanatory

variables for the half-year immediately ahead yields the corresponding forecast of enlistments. This procedure, of adding half-yearly observations, reestimating, and making a one-step ahead forecast, is repeated to derive our second set of forecasts.<sup>9</sup>

The performance of the two sets of forecasts is assessed by means of a number of standard techniques, most of which were originally proposed by Henri Theil (1966). Our measure of overall predictive accuracy is the standardized root-mean-square error, Theil's inequality coefficient  $U$ , given by

$$(3) \quad U = \left[ n^{-1} \sum (X_t - Y_t)^2 / n^{-1} \sum Y_t^2 \right]^{1/2},$$

where  $X$  and  $Y$  denote forecasts and outcomes, respectively, and  $t$  denotes time, there being  $n$  pairs of observations in the sample. The coefficient  $U$  is zero only in the extreme case of perfect forecasts, rises with inaccuracy, and has no upper bound.  $U$  equals 1 for any series of forecasts as inaccurate as a naive repetitive prediction of zero. The numerator of the inequality coefficient can be decomposed as follows:

$$(4) \quad \frac{1}{n} \sum (X_t - Y_t)^2 = (\bar{X} - \bar{Y})^2 + (s_x - rs_y)^2 + (1 - r^2)s_y^2,$$

where  $\bar{X}$  and  $\bar{Y}$  are the means, and  $s_x$  and  $s_y$  the standard deviations of forecasts and outcomes, respectively, and  $r$  is the correlation coefficient between the two series. Dividing both sides of (4) by the numerator of the inequality coefficient leads to the following three components:

$$(5) \quad UM = (\bar{X} - \bar{Y})^2 / \frac{1}{n} \sum (X_t - Y_t)^2$$

$$(6) \quad UR = (s_x - rs_y)^2 / \frac{1}{n} \sum (X_t - Y_t)^2$$

$$(7) \quad UD = (1 - r^2)s_y^2 / \frac{1}{n} \sum (X_t - Y_t)^2$$

<sup>9</sup>Ash (1980) argues that this sequential or recursive procedure is the correct way in which to estimate rational or other expectations functions, when information on expectations are not more directly observable, for example, from surveys.

TABLE 1—ACCURACY ANALYSIS OF FORECASTS OF THE ENLISTMENT RATE, 1977 (FIRST-HALF)—1979 (SECOND-HALF)

	Multiperiod Forecasts						One-Step-Ahead Forecasts					
	<i>U</i>	<i>UM</i>	<i>UR</i>	<i>UD</i>	<i>r</i>	<i>MAE%</i>	<i>U</i>	<i>UM</i>	<i>UR</i>	<i>UD</i>	<i>r</i>	<i>MAE%</i>
DOD												
All Males	.130	.224	.065	.711	.423	11.2	.140	.009	.256	.735	.074	12.3
White	.121	.079	.021	.900	.566	10.9	.143	.000	.153	.847	.324	12.4
Nonwhite	.100	.006	.121	.873	.190	8.1	.128	.225	.242	.533	.165	9.3
Army												
All Males	.137	.059	.034	.907	.467	11.3	.164	.014	.229	.757	.257	13.1
White	.153	.201	.000	.799	.656	12.7	.175	.025	.160	.816	.494	14.4
Nonwhite	.165	.268	.343	.389	-.203	15.1	.164	.128	.636	.236	-.650	12.2
Navy												
All Males	.176	.139	.015	.846	.350	13.8	.201	.095	.177	.728	.090	15.1
White	.210	.328	.024	.648	.409	16.6	.208	.151	.111	.738	.265	14.4
Nonwhite	.230	.452	.228	.320	.321	19.7	.223	.357	.263	.380	.042	15.9
Marines												
All Males	.225	.143	.226	.631	-.184	20.5	.243	.048	.733	.219	-.779 <sup>a</sup>	23.4
White	.202	.013	.162	.824	-.005	19.9	.256	.004	.746	.250	-.710 <sup>a</sup>	25.2
Nonwhite	.214	.249	.190	.562	.621	18.1	.204	.004	.047	.949	.248	19.0
Air Force												
All Males	.455	.863	.124	.013	.023	42.3	.171	.791	.176	.033	-.803 <sup>a</sup>	15.2
White	.423	.887	.096	.018	.322	39.9	.167	.732	.217	.050	-.776 <sup>a</sup>	14.4
Nonwhite	.366	.659	.323	.018	-.909 <sup>a</sup>	31.4	.154	.715	.042	.243	.768 <sup>a</sup>	14.7

Notes: *U*, *UM*, *UR*, *UD*, and *r* are defined in the text. *MAE%* denotes the mean absolute coefficient.

<sup>a</sup>Indicates a correlation coefficient significant at the 5 percent level.

where *UM* is the proportion of forecasting error (as measured by the numerator of *U*) due to bias; *UR* is the regression proportion or the proportion of error due to misforecasting the systematic component of the variance of outcomes; *UD* is the disturbance or residual proportion.<sup>10</sup> It would have been desirable to have carried out a full-time series analysis of the errors, but given the short sample period such an exercise would have been unlikely to yield any significant results. Indeed for this same reason whatever conclusions one draws from the accuracy analysis should be viewed at present as tentative.

<sup>10</sup>Consider the regression of outcomes on predictions,  $Y_t = \alpha + \beta X_t + \varepsilon_t$ , where  $\varepsilon_t$  is a random disturbance. For a series of optimal forecasts the following conditions hold:

$$\alpha = 0; \quad \bar{X} = \bar{Y}; \quad UM = 0$$

$$\beta = 1; \quad s_x = rs_y; \quad UR = 0.$$

Ideally then, because economic models are not deterministic, we would hope that *U* would be as small as possible while *UD* tended to unity.

Table 1 reports the results of the accuracy analysis of both multiperiod and one-step-ahead forecasts of the enlistment rate. Judged by the modest size of the inequality coefficients, the accuracy of the multiperiod forecasts is impressive. The mean value of *U* is 0.220, and excluding all forecasts for the Air Force, the least successful, this is reduced to 0.172. The mean absolute error (*MAE%*) averaged across the DOD totals, Army, Navy, and Marines, is a little under 15 percent of actual enlistments. Our most accurate forecasts are those for total DOD enlistments of nonwhites. There is no significant positive correlation between predictions and outcomes, not that this alone is much of a criterion for accuracy. Again ignoring the Air Force forecasts, systematic error is virtually absent: the relevant mean value for *UD* shows that approximately 70 percent of total forecasting error is random. There is some evidence of bias, namely, overestimating the average enlistment rate of nonwhites into the Army, Navy, and Marines, and of whites into the Navy, a tendency most apparent in the final two years of our chain of forecasts.

A negative correlation between forecasts and outcomes is responsible for the relatively high regression proportion,  $UR$ , in the forecasts of nonwhite Army enlistments, though this perverse correlation is not statistically significant at the 5 percent level.

Our least successful multiperiod forecasts are those for the Air Force. Here there is very marked downward bias, the enlistment rate underpredicted on average, resulting in unacceptable values for the mean absolute errors. During the period in question, the white enlistment rate fluctuated around a mean of approximately 8.5 percent, while the enlistment rate for nonwhites grew fairly steadily from about 6.5 to just under 9.5 percent of the relevant population. On the other hand, both our white and nonwhite equations lean heavily on a negative time trend, reducing the enlistment rate by some 1.5 percentage points a year, *ceteris paribus*, and hence the biased forecasts.

Fortunately, the predictive performance of the Air Force equations is much improved by sequential reestimation. Indeed, Table 1 shows that in terms of inequality coefficients and mean absolute errors the Air Force one-step-ahead forecasts are comparable to those for the Army and Navy, and substantially better than those for the Marines. Downward bias still persists even though reestimation reduces the absolute value of the equations' trend coefficient.<sup>11</sup> For the rest of the one-step-ahead forecasts, overall accuracy remains exceptionally good for the DOD totals, Army, and Navy. The accuracy of the Marine Corps' forecasts deteriorates somewhat: strong negative correlation between predictions and outcomes for whites gives rise to a high proportion of systematic error. The same problem is also apparent though to a lesser degree in the forecasts of nonwhite enlistments in the Army. Forecasts of nonwhite

naval enlistments tend to be biased upwards. Otherwise marked systematic error is absent.

With very few exceptions, our forecasts of the enlistment rate seem to be highly accurate. However, as C. W. J. Granger and P. Newbold point out:

It is well known... that the typical time series of economic levels is a near random walk. For such a series it is a very simple matter indeed to convince oneself that one has an excellent predictor of level. Indeed the simple 'no-change' appears very impressive in this light.  
[1973, p. 44]

They suggest that one's results are likely to be less flattering but more meaningful when one evaluates the accuracy of forecast changes. This is done in Table 2. The relevant variables are now the forecast and actual percentage changes in the enlistment rate, where the base upon which each percentage change is calculated is the enlistment rate in the previous half-year. In Table 2 the mean absolute error is replaced by the proportion of all forecasts which commit turning point errors ( $E$ ); that is, the proportion for which predicted and actual changes have opposite signs. For some purposes it may be more useful to forecast correctly the direction of change even at the expense of a larger absolute error than might be obtained with a forecast change that was closer numerically but wrong-signed.

Consider first the multiperiod forecasts. Not surprisingly, they now appear less accurate. Transforming the data into percentage changes raises the benchmark for success: the inequality coefficient is now unity for a naive no-change forecast, as against a naive no-enlistment forecast formerly. The mean value of  $U$  is 0.704 after excluding the outlying coefficients for the Air Force forecasts, which are markedly less accurate than assuming no change from one half-year to the next. Excluding the Air Force, all but one of the forecast series show significant positive correlation with the corresponding outcomes. No turning-point errors are committed by the forecasts for the Marines, even though this series fluctuates at least as much as any

<sup>11</sup>This finding, that aversion to military service is on the decline, is consistent with a recent survey of attitudes among Army officers and college students which concluded, *inter alia*, that "...1974-75 marked the beginning of a shift in American public opinion toward greater internationalism, support for defense, and concern about the Soviet Union" (Eugene Rosi, 1980, p. 16).

TABLE 2—ACCURACY ANALYSIS OF FORECASTS OF PERCENTAGE CHANGES IN THE ENLISTMENT RATE, 1977 (FIRST HALF)—1979 (SECOND-HALF)

	Multiperiod Forecasts						One-Step-Ahead Forecasts					
	U	UM	UR	UD	r	E	U	UM	UR	UD	r	E
DOD												
All Males	.760	.272	.328	.400	.875 <sup>a</sup>	.333	.754	.029	.134	.838	.720 <sup>a</sup>	.333
White	.664	.120	.330	.550	.866 <sup>a</sup>	.333	.743	.006	.121	.873	.708 <sup>a</sup>	.167
Nonwhite	.542	.001	.082	.917	.850 <sup>a</sup>	0	.672	.218	.018	.764	.802 <sup>a</sup>	.167
Army												
All Males	.698	.044	.139	.816	.768 <sup>a</sup>	.333	.823	.004	.029	.967	.570	.500
White	.747	.231	.002	.767	.731 <sup>a</sup>	.333	.825	.010	.013	.977	.525	.500
Nonwhite	.950	.272	.084	.644	.642	.167	.873	.125	.004	.872	.572	.167
Navy												
All Males	.576	.154	.023	.823	.852 <sup>a</sup>	.167	.646	.072	.104	.824	.809 <sup>a</sup>	.167
White	.717	.387	.001	.613	.825 <sup>a</sup>	.333	.664	.129	.123	.748	.816 <sup>a</sup>	0
Nonwhite	.767	.475	.038	.487	.839 <sup>a</sup>	.333	.706	.385	.001	.615	.827 <sup>a</sup>	.167
Marines												
All Males	.734	.214	.641	.146	.958 <sup>a</sup>	0	.744	.121	.705	.175	.948 <sup>a</sup>	.167
White	.615	.075	.882	.043	.992 <sup>a</sup>	0	.757	.050	.620	.330	.898 <sup>a</sup>	.167
Nonwhite	.682	.300	.575	.125	.966 <sup>a</sup>	0	.582	.042	.692	.266	.947 <sup>a</sup>	0
Air Force												
All Males	5.075	.838	.129	.034	-.361	.500	1.876	.778	.001	.221	.473	.500
White	4.620	.852	.109	.040	-.380	.500	1.768	.727	.012	.261	.419	.500
Nonwhite	3.104	.683	.247	.070	.112	.667	1.400	.691	.026	.283	.436	.667

Notes: E denotes the proportion of turning-point errors. (See text for definitions.)

<sup>a</sup>Indicates a correlation coefficient significant at the 5 percent level.

other; for the rest other than the Air Force, the worst misforecast one sign in three. Turning-point errors were often made when enlistments rose, in the second half of 1977, and when forecasting the decline and subsequent recovery which characterized 1979.

Nonsystematic error averages 43 percent of root mean square error overall, 53 percent excluding the Air Force, and 67 percent excluding both Air Force and Marines. Bias is most pronounced for the Air Force where, for reasons already indicated, the average growth of enlistments is consistently underpredicted. To a much smaller extent the forecasts of enlistments by whites and nonwhites into the Navy are characterized by overoptimistic bias. Systematic error in forecasting the variance of changes is proportionately high throughout the Marines' forecasts, and in the DOD totals for whites and all males; in all these cases the forecasts predict the systematic component of fluctuations too conservatively. Judged both by high accuracy and a low component of systematic error, our most successful forecasts are for total DOD enlistments by nonwhites.

Table 2 also shows the two principal advantages which accrue from updating the equation estimates in order to generate one-step-ahead forecasts. First, there is a marked improvement in the accuracy of the Air Force predictions: judged by the inequality coefficient, error is reduced by some 60 percent. (Nonetheless, they are still inferior to a naive no-change forecast, and the bias proportion remains high as does the proportion of turning-point errors.) Secondly, for all but one of the forecasts, the proportion of systematic error is smaller. Bias and regression proportions become very low for DOD, Army, and Navy forecasts. There remains some tendency to overpredict the growth of nonwhite enlistments in the Navy, to underpredict all enlistments in the Air Force, and to underpredict the systematic component of fluctuations in volunteers to the Marines.

On the whole the results of the accuracy analysis are encouraging. They show that economic theory can indeed enunciate a model capable of predicting future military enlistments with reasonable success (though we would be the first to recommend caution,

TABLE 3—MILITARY PERSONNEL SUPPLY FUNCTIONS, 1967 (SECOND-HALF)–1979(SECOND-HALF)

	Intercept	Pay Variable	Unemploy- ment Rate	Induction Variable	Time Trend	R <sup>2</sup>	DW	$\sigma$	Pay Elasticity	Unemploy- ment Elasticity
<b>DOD</b>										
<i>Enlistment Rate</i>										
All Males	1.340 (.191)	-.355 (.142)	-.200 (.569)	-.283 (.125)	-.014 (.003)	.834	2.170	.058	.655 .814	-
<i>Accession Rate</i>										
All Males	2.340 (.191)	-.355 (.142)	-.200 (.569)	-1.283 (.125)	-.014 (.003)	.967	2.170	.058	.538 .814	-
White	2.333 (.201)	-.342 (.149)	-.081 (.613)	-1.290 (.131)	-.017 (.003)	.968	2.172	.059	.538 .462	.015
Nonwhite	2.338 (.251)	-.779 (.355)	-.937 (.549)	-1.205 (.159)	.004 (.005)	.881	1.824	.078	.888 .459	-
<b>Army</b>										
<i>Enlistment Rate</i>										
All Males	.545 (.084)	-.196 (.060)	.215 (.257)	-.079 (.057)	-.006 (.001)	.714	2.541	.026	.881 1.088	.133
<i>Accession Rate</i>										
All Males	1.514 (.083)	-.197 (.059)	.219 (.253)	-1.042 (.056)	-.007 (.001)	.987	2.737	.026	.586 1.091	.090
White	1.463 (.081)	-.164 (.057)	.353 (.254)	-1.024 (.055)	-.009 (.001)	.989	2.747	.025	.529 1.217	.135
Nonwhite	1.826 (.159)	-.774 (.213)	-.216 (.354)	-1.169 (.103)	.005 (.003)	.919	1.991	.050	.763 .776	-
<b>Navy</b>										
<i>Enlistment Rate</i>										
All Males	.388 (.071)	-.111 (.055)	-.158 (.206)	-.113 (.045)	-.003 (.001)	.754	1.967	.021	.877 1.176	-
<i>Accession Rate</i>										
All Males	.388 (.071)	-.111 (.055)	-.158 (.206)	-.113 (.045)	-.003 (.001)	.754	1.967	.021	.877 1.176	-
White	.428 (.076)	-.131 (.059)	-.110 (.226)	-.127 (.048)	-.004 (.001)	.774	2.030	.022	.999 1.384	-
Nonwhite	.082 (.071)	.088 (.104)	-.155 (.152)	-.008 (.044)	.001 (.001)	.178	1.794	.021	- -	-
<b>Marine Corps</b>										
<i>Enlistment Rate</i>										
All Males	.210 (.037)	-.024 (.026)	-.075 (.114)	-.089 (.025)	-.002 (.001)	.854	2.543	.012	.314 .355	-
<i>Accession Rate</i>										
All Males	.242 (.038)	-.023 (.027)	-.079 (.116)	-.126 (.026)	-.001 (.001)	.889	2.286	.012	.296 .346	-
White	.241 (.039)	-.024 (.028)	-.023 (.123)	-.130 (.026)	-.002 (.001)	.896	2.286	.012	.328 .423	-
Nonwhite	.248 (.046)	-.015 (.062)	-.233 (.103)	-.114 (.030)	.001 (.001)	.728	2.051	.015	.072 .056	-
<b>Air Force</b>										
<i>Enlistment Rate</i>										
All Males	.188 (.062)	-.010 (.048)	-.186 (.179)	-.003 (.039)	-.003 (.001)	.679	1.682	.018	.088 .102	-
<i>Accession Rate</i>										
All Males	.188 (.062)	-.010 (.048)	-.186 (.179)	-.003 (.039)	-.003 (.001)	.679	1.682	.018	.088 .102	-
White	.205 (.062)	-.022 (.048)	-.151 (.186)	-.014 (.040)	-.003 (.001)	.692	1.697	.018	.190 .216	-
Nonwhite	.093 (.074)	.104 (.109)	-.224 (.159)	.056 (.046)	-.002 (.001)	.541	1.671	.023	- -	-

Notes: All equations estimated by two-stage least squares over sample period.

R<sup>2</sup> denotes square of multiple correlation coefficient.

DW denotes Durbin-Watson statistic.  $\sigma$  denotes standard error of regression. Standard errors of regression coefficients are shown in parentheses.

- denotes incorrectly signed coefficient.

Pay elasticities computed at both mean (first row) and terminal values (second row) of the relevant variables. Unemployment elasticities computed at the mean of the relevant variables.

again relying on the model's predictions for the Air Force). Whether or not these predictions are "good enough for the purpose in hand" (Milton Friedman, p. 41) can only be answered by those more intimately involved than ourselves in formulating and executing military manpower policy.

### III. The Model Reestimated

Encouraged by the model's track record, we conclude by presenting personnel supply functions reestimated over the full sample period running from the second half of 1967 through the second half of 1979. Table 3 shows coefficient estimates, elasticities, and equation summary statistics. Compared with our earlier results for the period 1968-76, almost all  $R^2$ 's are higher, and all equation standard errors are lower; although many Durbin-Watson values have drifted upwards, tests for negative first-order autocorrelation either refute the hypothesis or are inconclusive.

Present findings reinforce our earlier rejection of unemployment as a significant determinant of accessions or enlistments: only four of the unemployment coefficients have the correct, positive sign, and of these only one, for white Army accessions, exceeds its standard error. Total DOD enlistments increased by two for every seven inductees, the main beneficiary being the Navy, to which of course there was no direct draft. Negative trend coefficients are now smaller in absolute value for all armed forces except the Army; moreover the trend for nonwhite accessions is usually upwards, though not statistically significant.

There is no obvious general pattern when comparing old and new estimates of pay elasticities. The mean and terminal elasticities for total DOD enlistments, 0.655 and 0.814, respectively, are lower. On the other hand, all the Army pay elasticities and all the Navy's terminal pay elasticities are markedly higher, indeed they approach and in one case now exceed the estimated figure of 1.25 employed by the Gates Commission in their projection of accessions following the end of the draft (see the *Report*, President's Commission). Much lower pay elasticities are now

calculated for the Marine Corps and Air Force. The corresponding estimated coefficients are not significantly different from zero. We are not too surprised at these results for the Marines, a relatively small elite corps, perhaps the military service least amenable a priori to economic analysis. Besides, whatever reservations one might have concerning our analysis of enlistments in the Marines, the equations nonetheless perform adequately when used to generate out-of-sample forecasts. The Air Force equations are altogether less satisfactory. Demand constraints on Air Force enlistment may be tighter, while it is also well known that a high proportion of Air Force enlistees do so primarily in order to obtain on-the-job training in skills which will permit subsequent entry into similar civilian occupations. We are at present extending our research to incorporate both the possibility of demand constraint and the human capital aspect of the decision to enlist.

On the basis of our earlier results, we expressed concern that current military pay policy might be inadequate to meet manpower requirements in the 1980's. Were pay policy to be operated in a manner more competitive with the civilian labor market, our upward revision of some of the pay elasticities would point to a less gloomy future for the all-volunteer force; alternatively, should recent erosions in relative military pay continue, our former concern would only be deepened.

### REFERENCES

- Ash, Colin, "Irrational Estimation of Rational Expectations," Discussion Papers in Economics No. 150, University of Colorado, September 1980.
- Cooper, Richard V. L., *Military Manpower and the All Volunteer Force*, Santa Monica: The Rand Corporation, 1977.
- Fechter, Alan E., "The Supply of Enlisted Volunteers in the Post-Draft Environment: An Evaluation Based on Pre-1972," in Richard V. L. Cooper, ed., *Defense Manpower Policy: Presentations from the 1976 Rand Conference on Defense Manpower*, Santa Monica: The Rand Corpora-

- tion, 1979, 87-99.
- Fisher, Anthony C., "The Cost of the Draft and the Cost of Ending the Draft," *American Economic Review*, June 1969, 59, 239-54.
- Friedman, Milton, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.
- Granger, C. W. J. and P. Newbold, "Some Comments on the Evaluation of Forecasts," *Applied Economics*, March 1973, 5, 35-47.
- Grissmer, David W., "The Supply of Enlisted Volunteers in the Post-Draft Environment: An Analysis Based on Monthly Data, 1970-75," in Richard V. L. Cooper, ed., *Defense Manpower Policy: ...*, Santa Monica: The Rand Corporation, 1979, 100-15.
- McNown, Robert F., Bernard Udis, and Colin Ash, "Economic Analysis of the All-Volunteer Force," *Armed Forces and Society*, Fall 1980, 7, 113-32.
- Moses, Eleanor, "Is Equilibrium Just Equilibrium? A Critique of Non-Walrasian General Equilibrium Models," Discussion Papers in Economics Series A, No. 118, University of Reading, November 1980.
- Rosi, Eugene J., *Army Officers' Perceptions of International Politics: A Panel Study of the U.S. Army War College Class of 1975*, prepared for the 20th Anniversary Conference of the Inter-University Seminar on Armed Forces and Society, University of Chicago, October, 23-25, 1980.
- Theil, Henri, *Applied Economic Forecasting*, Amsterdam: North-Holland Publishers, 1966.
- Weintraub, E. Roy, *Microfoundations: The Compatibility of Microeconomics and Macroeconomics*, Cambridge: Cambridge University Press, 1979.
- Withers, Glenn, A., "International Comparisons in Manpower Supply," in Richard V. L. Cooper, ed., *Defense Manpower Policy: ...*, Santa Monica: The Rand Corporation, 1979, 116-36.
- President's Commission on an All-Volunteer Force, *The Report of the President's Commission on an All-Volunteer Force*, Washington: USGPO, February 1970.

## Working Capital Finance Considerations in National Income Theory

By DOUGLAS R. SHALLER\*

It is a fact of entrepreneurial life that production, production planning, and the sale and distribution of goods and services takes time. Ideally, all future contingencies could be resolved in advance of production by an appropriate set of contractual arrangements as in Arrow-Debreu general equilibrium models (see, for example, Gerard Debreu or Kenneth Arrow and Frank Hahn). However, as Edmond Malinvaud, James Tobin (1980), and others have argued, most contingent claims markets, including futures markets, do not exist.<sup>1</sup> Thus, firm managers must acquire stocks of financial assets and commodities in order to offset costs associated with the unsynchronized revenue and expenditure flows caused by time-consuming production and distribution processes.

From the economist's point of view, acquisition of stocks takes place until the marginal benefits of holding each asset in the portfolio just equals its marginal holding cost. For instance, in the case of money or commodities, the marginal benefits can come from lower transactions costs to the firm (William Baumol; Tobin, 1956; or Merton Miller and Daniel Orr). For customer loans, the benefits may arise due to lower transactions costs to the firm's customers (Arthur Okun; my thesis). In either case, the underlying unsynchronized cash flows must induce a

demand for the financial liabilities or equity capital necessary to finance both the fixed and working capital assets desired by the firm's managers.

Early concern over working capital finance can be found in the many writings of R. G. Hawtrey as well as in Oscar Lange's attempt to integrate interest into the theory of production.<sup>2</sup> Recently, Alan Blinder and Stanley Fischer have pursued the case of real interest rates negatively influencing *desired* inventories and thereby production in a *log* linear macromodel.

Concentrating on workers' behavior, Robert Lucas and Leonard Rapping assumed that the substitution effect between future and present leisure dominates the income effect. Thus, increases in the real interest rate cause increases in labor supply and, therefore, increases in the aggregate supply of output. However, they were not able to find empirical support for their hypothesized "Fisherian" interest rate effect on labor supply. Robert Barro (1980) presents an interesting model in the Lucas-Rapping tradition with rational expectations and a capital market where a positive Fisherian interest rate effect on aggregate supply plays a leading role.

But, while empirical evidence gathered so far does not seem to support a labor supply hypothesis which leads to a positive response of aggregate supply to real interest rate changes, there are indications that business behavior is influenced by working capital finance considerations so that there is, *ceteris paribus*, 1) a negative influence of real inter

\*Assistant professor of economics, Rutgers University. This article is based on a part of my dissertation. I would like to thank my thesis advisors, Hal Varian, Philip Howrey, and John Laitner.

<sup>1</sup>As Tobin recently put it,

True markets are rare and restricted in scope because the operation of such markets is expensive. The number of spot commodities in the U.S. economy is a large multiple of the human population. Rarely do two suppliers produce the same homogeneous commodity, and most firms sell an ever-changing menu of products. It is simply inconceivable that there could be organized competitive markets for them, let alone Walrasian multicommodity market clearing, that would dispense with the need for money. [1980, p. 89]

<sup>2</sup>Working capital in a long-period context is, of course, the hallmark of the Austrian school. Axel Leijonhufvud contains a brief but interesting introduction to the history of thought about capital theory. For a modern exposition of the Austrian approach, see C. C. von Weizsäcker.

est rates on output, and 2) a positive cost-push effect on supply price.

Recently, Leonard Sahling presented econometric evidence that indicates "that rising interest rates do exert a cost-push effect on prices, via the rental price of capital" (p. 924). Sahling estimates the long-run cost of capital elasticity of industrial prices to be .485 ( $t$ -value = 5.25), with nearly 80 percent of effect present within the first five quarters.

Multiple time-series analysis (on industrial production, money supply, commercial paper rate, producer price index) performed by Christopher Sims (1980) confirms a negative relationship between industrial production and surprise changes in short-term interest rates in aggregate U.S. time-series data. Sims discovers that 30 percent of the prediction variance of industrial production in the postwar period (16 percent in the interwar period) was accounted for by innovations in the four- to six-month prime commercial paper rate, while only 4 percent of variance (58 percent in the interwar period) was explained by surprises in the money supply ( $M1A$ —similar results hold for base money). Also, the possibility that working capital finance considerations have become more important in recent years is suggested by the fact that net customer credit per sales dollar has dramatically increased in the postwar period. For U.S. corporations, total accounts receivable minus total accounts payable all divided by sales increased nearly 80 percent from 1948 to 1970.<sup>3</sup>

The purpose of this paper is to explore working capital finance considerations in national income theory. For comparison, I analyze two macro models that reflect prominent but opposing strands in the macroeconomics literature. Interestingly, the one policy message that emerges in the context of these models is that tight money-loose fiscal policy regimes have singularly undesirable consequences tied to working capital finance effects.

In Section I, an analysis of the implications of the working capital finance hypothe-

sis for policy predictions in an equilibrium aggregate demand-aggregate supply model with price-taking firms is presented. The model in Section II is based on price-making firms and output adjustment to disequilibrium in the goods market. Section III is on the implications of the working capital finance hypothesis and the government budget constraint for bond-financed fiscal policy.

### I. The Aggregate Demand-Aggregate Supply Equilibrium Model

In this section, price-taking firms and workers exchange labor services for money at a market-clearing wage rate. Firms and consumers exchange output for money at a market-clearing price. An analysis of the implications of the working capital finance hypothesis follows the formal presentation of the model.

#### A. Working Capital Finance at the Firm Level

For an extremely simple example of working capital, suppose that inputs uniformly precede outputs by  $h$  units of time in a point-input, point-output production process so that the labor services  $n$  and a capital stock  $K$  that are entered into the firm's production process at time  $t$  are uniquely responsible for finished output at  $t + h$ . Thus work stays in process for  $h$  units of time and thereby must be financed.

In this simple example, optimal behavior for a competitive firm can be shown to yield an implicit short-period labor demand schedule equal to  $w = pf_n(n, K; h)\exp(-rh)$  where  $w$  is the nominal wage rate,  $p$  is the price level,  $f_n$  is the marginal product of labor,  $n$  is labor demand,  $K$  is the capital stock,  $h$  is the production lag, and  $r$  is the anticipated real rate of interest.

In other words, when the firm chooses inputs, optimal behavior with nonsynchronized cash flows requires that a discounted marginal revenue product be associated with a specific marginal factor cost. The necessity of this present value calculation stems from the opportunity cost of funds tied up in

<sup>3</sup> Calculated from data contained in *Historical Statistics of the U.S., Colonial Times to 1970*.

working capital. Other examples of working capital that yield results similar to the work-in-process example described above include finished goods and raw materials inventories and accounts receivable (customer loans).

### B. An Equilibrium Macro Model with Working Capital

Let the supply-side equilibrium (aggregate supply) levels of employment  $n$  and output  $y$  be defined by

$$(1) \quad (p^*/p)g(n) = f_n(n, K; h) \exp(-rh) \\ g' > 0$$

$$y = f(n, K; h),$$

where  $p^*/p$  is the ratio of the price level anticipated by workers to the actual price level and  $g(n)$  is a neoclassical labor supply price curve based on a correctly anticipated price level. Thus  $(p^*/p)g(n)$  is the actual supply price of workers in real terms. Also,  $f(n, K; h)$  is the production function previously discussed. Clearly, under the usual regularity conditions, we can collapse the two-equation system above into a short-period aggregate supply function that is sensitive to the cost of working capital finance. Formally we can write the aggregate supply function as<sup>4</sup>

$$(2) \quad y = y(r, p/p^*),$$

with  $y_r < 0$  and  $y_{p/p^*} > 0$ . In the long run, the aggregate supply function becomes  $y(r)$  as  $p^*$  converges to  $p$ . Moreover, with perfect foresight aggregate supply is  $y(r)$ .

Suppose that aggregate demand can be represented by the familiar conditions of flow equilibrium in the goods market,

$$(3) \quad y = C(y - T, W) + I(r) + G,$$

with  $C_{y-T} > 0$ ,  $C_W > 0$ , and  $I_r < 0$ , where  $C(y - T, W)$  is the consumption function,  $I(r)$  a disequilibrium demand for invest-

ment,<sup>5</sup>  $T$  is taxes,  $W$  is wealth, and  $G$  is government spending. Furthermore, for expository simplicity, assume that taxes are exogenous, wealth is not an argument of the money demand function,<sup>6</sup> and individuals value real capital at replacement cost,<sup>7</sup> so that stock equilibrium in the money market is represented by

$$(4) \quad M/p = L(y, r + \pi),$$

with  $L_y > 0$ ,  $L_{r+\pi} < 0$ , and where  $M$  is the stock of nominal government fiat money,  $L(y, r + \pi)$  is money demand, and  $\pi$  is the anticipated rate of inflation. Also, real wealth is defined by

$$(5) \quad W = K + M/p + B/p,$$

where  $B$  is the stock of government bonds outstanding.<sup>8</sup>

<sup>5</sup>As is usual for short-run macro theory, by assumption, there does not exist a market for used capital goods. Therefore, firms must accumulate or decumulate current output in order to change their capital stock. The disequilibrium investment demand that results when the existing capital stock cannot be instantaneously adjusted to its desired level. Furthermore, all investment is financed by issuing shares and all profits are returned to shareholders. Thus, the disequilibrium demand for investment is a function of the real cost of finance, the instantaneous real rate of return on shares which must equal the instantaneous real rate of return on alternative interest bearing securities given arbitrage, perfect certainty, and no tax differences.

<sup>6</sup>Here I am assuming that  $L_W = 0$ . Therefore, we must have  $B_W^d = 1$  where  $B^d$  is the demand for bonds. This assumption is not crucial for the short-run results presented herein. Moreover, the assumption is relaxed in the latter section that deals with the government budget constraint.

<sup>7</sup>Thus, individuals ignore the present value of scarcity rents in their evaluation of net worth. This simplifying assumption is quite common. Furthermore, relaxing this assumption does not significantly alter my major conclusions.

<sup>8</sup>In this model government bonds are the only alternative to holding shares other than non-interest-bearing government fiat money. Therefore  $r$  is both the real rate of return on shares and the real return on government bonds. The government bonds are assumed to be very short so that the capital gains and losses due to interest rate changes are nonexistent. This assumption is for expository simplicity, only and does not qualitatively affect the results.

<sup>4</sup>Subscripts denote partial derivatives.

In a model with price-taking agents, it is conventional to specify Walrasian dynamic assumptions. For convenience sake, as will become clear below, I will do likewise and suppose that prices and interest rates move according to Walras' excess demand hypothesis so that

$$(6) \quad \dot{p} = s_1 [C(y - T, W) + I(r) + G - y(r, p/p^*)]$$

$$\dot{r} = s_2 [L(y, r + \pi) - M/p]$$

and  $s_1[0] = s_2[0] = 0$ ,  $s'_1 > 0$ ,  $s'_2 > 0$ .

### C. Policy Experiments

Totally differentiating (6) at the presumed unique point of "temporary" equilibrium where  $p^*$  and  $\pi$  are constant and  $\dot{p} = \dot{r} = 0$ , we can formally uncover the short-run (in this model, unanticipated) comparative static results of interest. Interestingly, the working capital finance effect  $y_r < 0$  renders the fundamental determinant of the system indeterminate as to sign. The fundamental determinant can be written

$$(7) \quad D = \alpha\delta - \beta\gamma$$

$$\text{where } \alpha = -((1 - C_y)y_p)/p^* - C_w(M + B)/p^2 < 0,$$

$$\delta = L_y y_r + L_r < 0,$$

$$\beta = -(I - C_y)y_r + I_r = ?,$$

since the first term is positive and the second negative, and

$$\gamma = (L_y y_p)/p^* + M/p^2 > 0.$$

Therefore, all comparative statics derivatives are indeterminate.

However, when the working capital finance effect  $y_r < 0$  is not overpowering, the sensitivity of aggregate supply to real interest

rate changes is less than that of aggregate demand. Moreover, it is a straightforward task to show that this case is equivalent to the determinantal stability condition,  $D > 0$ . Henceforth, I will simply assume that  $D > 0$ .

Thus, even when the aggregate supply curve is steeper than the aggregate demand curve, the unanticipated government spending multipliers on prices  $dp/dG > 0$  and interest  $dr/dG > 0$  are unambiguous, but the effect on income is indeterminate, since

$$(8) \quad dy/dG = (-L_r y_p + y_r M/p^2)/D.$$

with  $-L_r y_p > 0$ ,  $y_r M/p^2 < 0$ , and  $D > 0$ . The interpretation of (8) is clear cut. While the usual short-run supply effects raise income, the increase in the user cost of working capital lowers it.

This is a new and qualitatively different twist on the crowding-out theme.<sup>9</sup> Moreover, a necessary and sufficient condition of expansionary fiscal policy derived from the numerator of (8) is

$$(9) \quad -L_r/(M/p^2) + y_r/y_p > 0.$$

Multiplying both sides by  $r/p$  and rearranging, we find that expansionary fiscal policy requires that in terms of elasticity,<sup>10</sup>

$$(10) \quad E(L)/E(r) > E(p)/E(r).$$

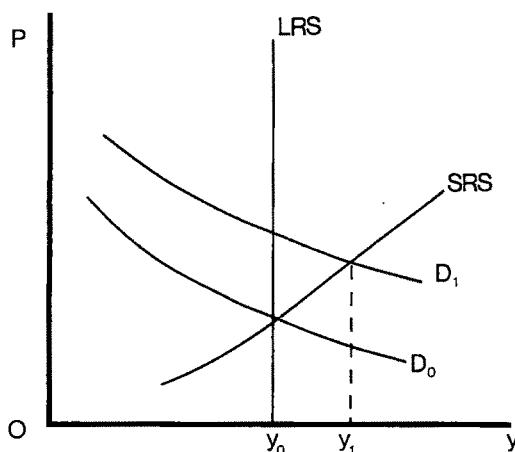
Hence, only when the interest elasticity of money demand exceeds the interest elasticity of supply price for final goods is unanticipated government spending expansive. In contrast, a fully anticipated ( $p = p^*$ ) increase in government spending implies that

$$(11) \quad dy/dG = y_r(M/p^2)/D < 0.$$

Furthermore, a fully anticipated balanced budget decrease in government spending and

<sup>9</sup>Notice that this result does not depend on Pigouvian wealth effects in the consumption function.

<sup>10</sup>In this paper,  $E(x)/E(y) = d \ln x / d \ln y$ . See Fischer for example.

FIGURE 1. EXPANSIONARY FISCAL POLICY: [ $y_R = 0$ ]

taxes yields

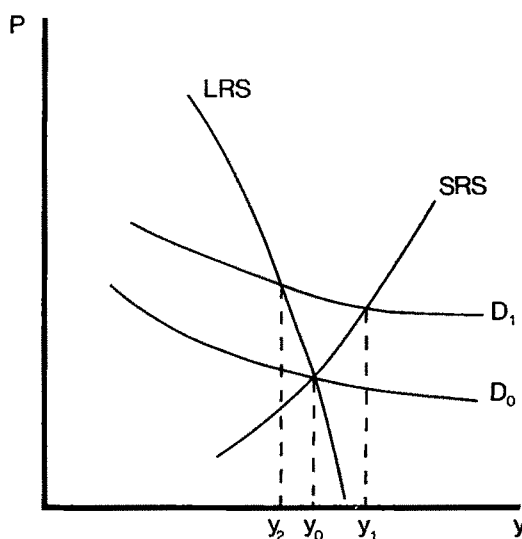
$$(12) \quad dy/dG^* = y_r(M/p^2)(1 - C_y)/D < 0.$$

Thus an anticipated reduction in government spending matched dollar for dollar by an anticipated tax reduction is in the long-run expansionary; but on the other hand, a straight income tax reduction lowers equilibrium national income.

To see this graphically, notice that the *LM* curve, equation (4), can be solved for *r* and then substituted into the aggregate supply function (2) as well as into the *IS* curve (3), to yield quasi-reduced-form aggregate supply and aggregate demand curves in (*P*, *y*) space.

Figures 1, 2, and 3 illustrate three possibilities. Figure 1 represents a surprise increase in government spending (a shift in the aggregate demand curve (*D*<sub>0</sub>)) when working capital finance effect on aggregate supply is ignored, ( $y_r = 0$ ). Starting from the natural or long-run rate of output  $y_0$ , output increases initially by  $y_1 - y_0$  along the short-run supply curve (*SRS*), but then recedes to its natural level at the long-run supply curve (*LRS*) when price expectations catch up to prices.

In Figures 2 and 3, the working capital finance hypothesis ( $y_r < 0$ ) means that the long-run aggregate supply curve (*LRS*) along with the aggregate demand curve (*D*<sub>0</sub>) is negatively sloped because a high price level

FIGURE 2. EXPANSIONARY FISCAL POLICY:  
[ $y_R < 0$ ,  $E(L) - E(R) > E(p) - E(R)$ ]

is, *ceteris paribus*, associated with a small real money supply and thus a high real interest rate. High real interest rates mean costly working capital finance thereby implying less aggregate supply in the same way that the high real interest rates translate into expensive fixed investment thereby implying less aggregate demand.

Figure 2 illustrates the case where the economic-agents-being-fooled effect outweighs the working-capital-finance-crowding-out effect due to higher real interest rates so that the short-run aggregate supply curve *SRS* is positively sloped. In this case, output first increases from  $y_0$  to  $y_1$ . But when expectations catch up, output decreases until it ends up at  $y_2$ , a lower level than where it began. Figure 3 illustrates the case where output decreases in the short run and in the long run because the temporary expansionary effect of the signal extraction problem is not powerful enough to offset the contractionary effect of fiscal finance.

In Figures 2 and 3, a fully anticipated increase in government spending or decrease in taxes is contractionary. However, in the case of a tax cut, if the substitution effect of a wage tax rate change dominates the income effect for labor supply, then the aggregate

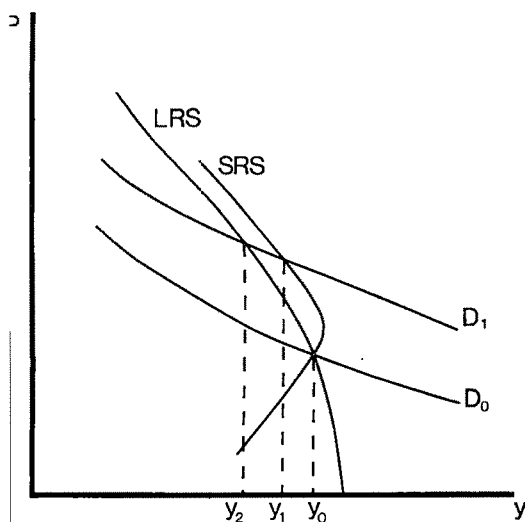


FIGURE 3. EXPANSIONARY FISCAL POLICY:  
 $[y_R < 0, E(L) - E(R) < E(P) - E(R)]$

supply curve would shift forward and the final result would be ambiguous. (See my forthcoming article for an analysis of tax incentive effects on labor supply in the context of a classical macro model.)

In summation, the working capital finance hypothesis means that fiscal finance raises the average and marginal costs of doing everyday business. Higher real interest rates not only make fixed capital finance more expensive and thus reduce business demand for investment goods, but also make working capital finance more costly, thereby reducing aggregate supply and equilibrium national product. In contrast, without a working capital finance effect, higher real interest rates merely transfer real product from private sector to public sector consumption.<sup>11</sup>

The comparative statics results for unanticipated monetary policy are

$$(13) \quad \begin{aligned} & (?) \\ dr/dM = & ((-y_p(1 - C_y - C_w L_y)) / p^* \\ & (+) \\ & - C_w(B/p^2)) / pD, \end{aligned}$$

<sup>11</sup>Of course, if the fiscal policy-induced significant substitution effects or changes in the distribution of wealth, then this result may not necessarily hold.

$$\begin{aligned} & (?) \\ dp/dM = & ((y_r(1 - C_y - C_w L_y) \\ & (-) \\ & - (I_r + C_w L_r)) / pD, \end{aligned}$$

and therefore

$$(14) \quad \begin{aligned} dy/dM = & ((-y_p(I_r + C_w L_r)) / p^* \\ & - y_r C_w(B/p^2)) / pD > 0. \end{aligned}$$

Clearly, with a small wealth effect on consumption, then  $dr/dM < 0$ . Notice, however, that in the ordinary case just described, more money pushes up prices by shifting aggregate demand, but pushes down prices by lowering costs. At any rate, the working capital finance effect quantitatively increases the effects of unanticipated money.

Furthermore, a fully anticipated decrease in the money stock lowers the equilibrium level of output and employment in the absence of wage or price or expectational rigidities when there are wealth effects in the consumption function, working capital finance effects in aggregate supply, and some proportion of government bonds are considered wealth by households.<sup>12</sup>

## II. Working Capital Finance in a Model with Output Constrained Price-Setting Firms

In this section, price-making sales expectations constrained firms are substituted for the price takers above. Suppose then that output responds to disequilibrium in the goods market and that the cost of working capital as well as the wage rate  $w$  enters into the determination of the supply price of output. Also suppose that the nominal wage rate is predetermined so that short-run aggregate supply responds negatively to downward shifts of the aggregate demand curve.<sup>13</sup>

<sup>12</sup>Under a similar set of assumptions, Blinder and Fischer demonstrate that money is not neutral when desired inventories are sensitive to the real interest rate. In this model, the carrying costs of existing work-in-process is the key linkage.

<sup>13</sup>See Fischer for a discrete time model with long-term wage contracts which provide for temporary wage rigidity.

Formally, let

$$(15) \quad \dot{y} = s_1 [C(y - T, W) + I(r) + G - y]$$

$$\dot{r} = s_2 [L(y, r + \pi) - M/p],$$

and

$$s_1[0] = s_2[0], \quad s'_1 > 0, \quad s_2 > 0;$$

where

$$(16) \quad p = p(y, r, w) \quad \text{and} \quad p_y, p_r, p_w > 0.$$

By taking the total differential of the system (15) at  $\dot{y} = \dot{r} = 0$ , it follows that the trace condition for local dynamic stability can be written

$$(17) \quad TR^* = (s'_1/s'_2)\alpha^* + \delta^* < 0$$

where

$$\alpha^* = -(1 - C_y) - p_y C_w (M + B)/p^2 < 0$$

$$\delta^* = L_r + p_r M/p^2 = ?$$

But when the interest rate moves much more quickly than output in response to disequilibrium, the trace condition in the limit can be interpreted as saying  $\delta^* > 0$ , or in other words,

$$(18) \quad E(L)/E(r) > E(p)/E(r),$$

which implies that  $dy/dG > 0$ . Therefore in this case since  $E(L)/E(r) < E(p)/E(r)$  implies  $dy/dG < 0$ ,  $dy/dG < 0$  characterizes an unstable equilibrium.

In contrast, under the neoclassical price-taking assumptions above, the trace condition is

$$(19) \quad TR = s'_1/s'_2 \alpha + \delta < 0,$$

$$\text{where} \quad \alpha = -(1 - C_y)y_p$$

$$- C_w (M + B)/p^2 < 0$$

$$\delta = L_r + y_r L_y < 0.$$

Therefore, the presumption of relatively

speedy adjustment in the money market yields in the limit

$$(20) \quad L_r + y_r L_y < 0,$$

or

$$(21) \quad E(L)/E(r) > y_r L_y r / (M/p^2) < 0,$$

which places no additional restrictions on the parameters of the system. Therefore, I cannot establish the stability of any full crowding-out short-run equilibria in terms of the usual conditions without a more fully specified model when firms set prices and perceive sales constraints. That is, in the case considered, when  $E(L)/E(r) < E(p)/E(r)$  so that  $dy/dG < 0$ , the usual self-correcting mechanisms in the model are not sufficiently powerful. On the other hand, some full crowding-out short-run equilibria are stable in the comparably specified neoclassical model of Section I.

Interestingly, it is reasonable to suppose both that  $E(L)/E(r)$  will become smaller as interest rates rise due to diminishing returns to economizing on money balances, and that  $E(p)/E(r)$  will not shrink but will perhaps increase as interest rates rise given economically appealing pricing procedures. Thus, it follows that since a loose fiscal policy and a tight monetary policy both raise real and nominal interest rates in these models, then the combination of a loose fiscal policy coupled with a tight monetary policy would be destabilizing in the sales-constrained model of this section when the policies are executed with sufficient resolution. In the neoclassical model of Section I, the same loose money-tight fiscal program would also reduce *GNP*, but would not necessarily be destabilizing. Of course, the usual caveats certainly apply to the interpretation of results obtained from mechanical models encompassing quite a number of maintained hypotheses.

### III. The Government Budget Constraint and the Crowding Out of Bond-Financed Government Spending

It is typically argued that while increased wealth, due to bond finance, increases consumption and shifts the *IS* curve out, the

larger real wealth, increases the demand for real balances, in effect shifting the *LM* curve back.<sup>14</sup> The net effect of changes in wealth therefore appears to be indeterminate, and has become the target of much debate between monetarists and fiscalists.

Milton Friedman has, for example, in recent years (1970, 1972), pinned his attack on the efficacy of pure fiscal policy on the presumed strength of, on balance, contractionary wealth effects, pointing to the fact that a deficit must be financed in all future time periods (the government budget constraint).

However, Blinder and Robert Solow (1973) obtained the striking result that contractionary net wealth effects are inconsistent with convergence to a balanced budget equilibrium. That is, an increase in government bonds, starting at a balanced budget equilibrium, will eventually cause the system to converge to a new balanced budget equilibrium only if the net effect of increased bonds is expansionary.

This result is not only devastating to Friedman's attack within the context of a standard model at the usual level of aggregation, but, also curious, since many economists argue that these secondary wealth effects are probably, on balance, contractionary. (See Keith Carlson and Roger Spencer, 1975.)

While the Blinder-Solow result is obtained under the assumption of fixed wages and prices, they maintain that the result goes through with flexible prices. However, under the more detailed supply-side consideration of this paper, an elasticity condition emerges.

Following Blinder and Solow, let us suppose that government bonds are fixed coupon perpetuities that pay one dollar per unit of time forever. Then

$$(22) \quad W = K + M/p + B/rp,$$

where *B* is both the nominal debt service and the number of government bonds outstand-

ing, so that *B/rp* is the real value of the outstanding debt.

Moreover, again following Blinder and Solow, assume an endogenous tax function of the form

$$(23) \quad T = T(y + B/p),$$

where  $T(0) = 0$  and  $T' > 0$ , so that disposable income is

$$(24) \quad y^d = y + B/p - T(y + B/p).$$

Therefore, the government budget constraint is

$$(25) \quad \dot{M}/p + \dot{B}/rp = G + B/p - T(y + B/p).$$

As has been argued, the government budget constraint serves only to enforce degrees-of-freedom consistency. Clearly, aggregate demand with aggregate supply forms a recursive subset of the total system, since the aggregate demand, aggregate supply subset or determines *y*, *r*, and *p*, conditional on *G*, *B*, and *M*.

However, to maintain a continuity of notation with the Blinder and Solow paper, let us use reduced-form equations for *r* and *y*, but keep the structural equation for *p*. Thus, let the recursive block be represented by

$$(26) \quad y = F(G, B, M); \quad r = H(G, B, M); \\ p = p(y, r).$$

In order to see what restrictions balanced budget stability puts on the multipliers of the system under bond financing, I will substitute the recursive block of equations (26) into the government budget constraint (25) and set  $\dot{M} = 0$ , leaving

$$(27) \quad \dot{B} = H(G, B, M)p(y, r) \\ \times \left( G + \frac{B}{p(y, r)} - T\left(y + \frac{B}{p(y, r)}\right) \right).$$

By straightforward logarithmic partial differ-

<sup>14</sup>Barro (1974) argues that government bonds are not net wealth when the present value of the future tax liabilities of future generations are considered by the current generation of households. However, this assumption implies a set of internalized social values that are quite special.

entiation and evaluation at equilibrium, we are left with

$$\begin{aligned}
 (28) \quad \dot{B}_B &= rp \left( (1 - T') \frac{p - B(p_y F_B + p_r H_B)}{p^2} \right. \\
 &\quad \left. - T' F_B \right) \\
 &= r \left( (1 - T') \frac{B}{p} p_y + p T' \right) F_B \\
 &\quad + r(1 - T') \frac{(p - B p_r H_B)}{p}.
 \end{aligned}$$

Thus assuming local dynamic stability around a balanced-budget equilibrium, we must have

$$(29) \quad dy/dB = F_B > \Omega$$

where

$$\Omega = (p - B p_r H_B) / (p_y + p^2 T' / (1 - T)),$$

with the denominator strictly positive and the numerator unsigned. Therefore, on balance, contractionary wealth effects ( $F_B < 0$ ) are consistent with balanced budget stability when

$$(30) \quad p - B p_r H_B < 0.$$

Notice that (30) can be rearranged and written in elasticity terms so that it becomes

$$(31) \quad E(p)/E(B) > 1.$$

The intuition behind this result is compelling. If the increase in bonds used to finance spending pushes prices up by a greater percentage than the percentage increase in bonds, then the real debt service shrinks. Moreover, a smaller real debt service means a smaller real budget deficit since the real debt service expenditure will shrink more than the real tax revenue on that real debt service.

In this case, when  $E(p)/E(B) > 1$ , the deficit will close over time even though the secondary wealth effects on consumption and

money demand, on balance, reduce *GNP*. Thus contractionary wealth effects can be consistent with convergence to a balanced budget equilibrium.

#### IV. Conclusions

My principal conclusions on the implications of working capital finance for national income theory are as follows: First, in an otherwise fairly standard neoclassical aggregate demand-aggregate supply model with price-taking firms and workers, (a) a fully anticipated "expansionary" fiscal policy in the absence of an accommodating monetary policy reduces real *GNP*, (b) an unanticipated policy may reduce real *GNP*, and (c) given wealth effects on consumption, a fully anticipated decrease in the money stock implies smaller equilibrium levels of output and employment. (This result does not depend on wage rigidities or money illusion.)

Second, in an aggregate demand-aggregate supply model with price-making firms and output adjustments to disequilibrium in the goods market; a) "full crowding-out" short-run equilibria that result from unanticipated expansionary fiscal policies are possible but unstable, and b) loose fiscal-tight money regimes both reduce *GNP* and are unstable when executed with sufficient resolution.

Third, government budget constraint stability considerations yield neither necessary nor sufficient conditions that enable one to rule out intermediate-run crowding out in a stable model.

#### REFERENCES

- Arrow, Kenneth J. and Hahn, Frank H., *General Competitive Analysis*, San Francisco: Holden Day, 1971.
- Barro, Robert, "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, 82, 1095-117.
- \_\_\_\_\_, "A Capital Market in an Equilibrium Business Cycle Model," *Econometrica*, September 1980, 48, 1393-418.
- Baumol, William J., "The Transaction Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics*, November 1952, 66, 545-56.

- Bлиндер, Alan S. and Fischer, Stanley, "Inventories, Rational Expectations, and the Business Cycle," Working Paper No. 381, National Bureau of Economic Research, 1979.
- \_\_\_\_\_ and Solow, Robert M., "Does Fiscal Policy Matter?," *Journal of Public Economics*, November 1973, 2, 318-37.
- Carlson, Keith M. and Spencer, Roger W., "Crowding Out and Its Critics," *Federal Reserve Bank of St. Louis Review*, December 1975, 57, 2-16.
- Debreu, Gerard, *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, New Haven: Yale University Press, 1959.
- Fischer, Stanley, "Long-Term Contracts, Rational Expectations and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, 85, 191-205.
- Friedman, Milton, "The Role of Monetary Policy," *American Economic Review*, March 1968, 58, 1-17.
- \_\_\_\_\_, "A Theoretical Framework for Monetary Analysis," *Journal of Political Economy*, March/April 1970, 78, 193-238.
- \_\_\_\_\_, "Comments on the Critics," *Journal of Political Economy*, September/October 1972, 80, 906-50.
- Hawtrey, R. G., *Trade and Credit*, London: Longmans, Green & Co, 1928.
- \_\_\_\_\_, *Capital and Employment*, London: Longmans, Green & Co, 1937.
- Holt, Charles C. et al., *Planning Production, Inventories, and Work Force*, Englewood Cliffs, New Jersey: Prentice-Hall, 1960.
- Lange, Oscar, "The Place of Interest in the Theory of Production," *Review of Economic Studies*, June 1936, 3, 159-92.
- Leijonhufvud, Axel, *On Keynesian Economics and the Economics of Keynes*, London: Oxford University Press, 1968.
- Lucas, Robert E. Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Alan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19-46.
- \_\_\_\_\_ and Rapping, Leonard A., "Real Wages, Employment, and Inflation," *Journal of Political Economy*, September/October 1969, 77, 721-54.
- Malinvaud, Edmond, "Interest Rates in the Allocation of Resources," in F. H. Hahn and F. P. R. Brechling, eds., *The Theory of Interest Rates*, London: Macmillan Co., 1965, 217-18.
- Miller, Merton H. and Orr, Daniel, "A Model of the Demand for Money by Firms," *Quarterly Journal of Economics*, August 1966, 80, 413-35.
- Okun, Arthur M., *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.
- Sahling, Leonard, "Price Behavior in U.S. Manufacturing: An Empirical Analysis of the Speed of Adjustment," *American Economic Review*, December 1977, 70, 911-25.
- Shaller, Douglas R., "Financial Linkages on the Supply Side: Three Essays on Business Behavior, 'Working Capital' Finance, and National Income Theory," unpublished doctoral dissertation, University of Michigan, 1980.
- \_\_\_\_\_, "The Tax-Cut-But-Revenue-Will-Not-Decline Hypothesis and the Classical Macromodel," *Southern Economic Journal*, forthcoming.
- Sims, Christopher A., "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review Proceedings*, May 1980, 70, 250-57.
- Tobin, James, "The Interest-Elasticity of the Transactions Demand for Cash," *Review of Economics and Statistics*, August 1956, 38, 241-47.
- \_\_\_\_\_, "Discussion" in John H. Karenken and Neil Wallace eds., *Models of Monetary Economies, Proceedings and Contributions from Participants of a December 1978 Conference Sponsored by the Federal Reserve Bank of Minneapolis*, Minneapolis: Federal Reserve Bank of Minneapolis, 1980, 89.
- von Weizaecker, C. C. *Steady State Capital Theory*, New York: Springer-Verlag, 1971.
- U.S. Department of Commerce, Bureau of Census, *Historical Statistics of the United States, Colonial Times to 1970, Part 2*, Washington: USGPO, September 1975, 924-25.

# Substitution Between Wage and Nonwage Benefits

By STEPHEN A. WOODBURY\*

You gotta be kidding, Ollie. You're just robbing yourself. Schuylkill Mutual offers a terrific deal on Keogh, and we could plug you in, in fact we advise plugging you in, on the corporate end so not a nickel comes out of your personal pocket, it comes out of the corporate pocket and there's that much less for Uncle to tax. These poor saps carrying their own premiums with no company input are living in the dark ages. There's nothing shady about rigging it this way, we're just using the laws the government has put there. They *want* people to take advantage, it all works to up the gross national product. You know what I mean by Keogh, don't you? You're looking kind of blank.

*John Updike, p. 296*

The study of how labor is compensated, whether by wage or nonwage benefits, has taken on increasing importance as the proportion of total compensation that workers receive as wages and salary has declined. The interest of economists, labor relations experts, and policymakers in this shift has centered on six issues. First, there is the theoretical problem of knowing how well fringe benefits substitute for wage benefits in workers' preference patterns. The equalizing differences hypothesis—that other things equal, a desirable job characteristic can compensate for an undesirable job characteristic, pecuniary or nonpecuniary—has received a flurry of recent attention from economists, although only sporadic evidence in support of the hypothesis has been found.<sup>1</sup> Second,

the problem of measuring total real compensation has increased as a smaller fraction of compensation has been taken in an easily measured form such as wages. There is need to understand what sorts of bias are introduced by exclusive reliance on wages as a measure of compensation. Third, in discussions of collective bargaining in the public sector, it has been alleged that liberal fringe benefits received by public employees are jeopardizing the financial soundness of local governments. The veracity of such allegations hinges on whether or not lower fringe benefits necessitate higher wages in order to attract and retain public employees of given qualifications (see Ronald Ehrenberg). Fourth, the impact on employment costs of pension reform legislation such as the Employee Retirement Income Security Act (ERISA) depends crucially on the substitution possibilities that exist between wages and fringes. Fifth, the financial stability of the Social Security system depends on the degree to which that program's tax base erodes as the mix of compensation shifts away from wages and toward fringes (see Yung-Ping Chen). Finally, a host of practical questions relating to rational personnel management and planning are connected to workers' preferences for wages and benefits: What sort of fringe-wage "package" should be devised in order to induce workers of certain characteristics into the labor force? How much do young people entering the labor force care about fringe benefits? (We wish to find out by observing behavior, not by asking people.)

Table 1 displays summary statistics on the growth of fringe benefits during roughly the last fifteen years. Employees of firms surveyed by the Bureau of Labor Statistics (BLS) received 4.9 percent of their compensation in the form of supplements to wages and salaries in 1966, but that figure had grown to 9.2 percent by 1976. Supplements here refer to all voluntary expenditures made by the employer for compensation other than direct

\*Assistant professor of economics, Michigan State University. This work was supported in part by U.S. Department of Labor Grant No. 91-55-79-14. I am grateful for the advice and comments of Laurits R. Christensen, Alan L. Cohen, Sheldon Danziger, Daniel S. Hamermesh, W. Lee Hansen, Susan Pozo, Joseph F. Quinn, Mark J. Roberts, James L. Stern, and an anonymous reviewer.

<sup>1</sup>For excellent reviews, see Charles Brown, and Robert Smith.

TABLE 1—TRENDS IN WAGE AND NONWAGE COMPENSATION: 1965–78

Year	Compensation of Workers in Firms Surveyed by BLS <sup>a</sup>			Compensation of Workers in Firms Surveyed by Chamber of Commerce <sup>b</sup>		
	\$ per Hour <sup>c</sup> (Average)	Gross Payroll <sup>d</sup> (Percent of Total)	Supplements <sup>e</sup> (Percent of Total)	\$ per Hour <sup>c</sup> (Average)	Gross Payroll <sup>d</sup> (Percent of Total)	Supplements <sup>e</sup> (Percent of Total)
1965				\$2.99	90.4	9.6
1966	\$3.26	95.1	4.9			
1967				3.27	89.8	10.2
1968	3.70	94.6	5.4			
1969				3.59	89.1	10.9
1970	4.31	94.0	6.0			
1971				4.05	87.8	12.2
1972	4.94	93.2	6.8			
1973				4.61	87.0	13.0
1974	5.93	92.4	7.6			
1975				5.43	85.6	14.4
1976	7.05	90.8	9.2			
1977				6.01	84.2	15.8
1978				6.66	83.9	16.1

<sup>a</sup>Source: U.S. Department of Labor, Bureau of Labor Statistics.

<sup>b</sup>Source: Chamber of Commerce of the United States, 1978. (Constructed from data in Table 19.)

<sup>c</sup>Total compensation of employees excluding legally required payments such as Social Security, worker's compensation, and unemployment insurance.

<sup>d</sup>All direct payments to workers, including straight-time and premium-time pay, pay for time not worked (vacations, holidays, sick and other leave, and—in the Chamber of Commerce survey—pay for rest periods and lunch breaks), and nonproduction bonuses.

<sup>e</sup>Employer payments to pension, health and life insurance, and other agreed-upon items.

payments to workers. They thus include payments to private retirement systems as well as to life insurance, health benefits, and other agreed-upon plans. Excluded are legally required payments such as Social Security, worker's compensation, and unemployment insurance contributions. Figures gathered by the Chamber of Commerce of the United States show a similar growth of the percentage of compensation received as supplements. The 1965 figure was 9.6 percent, the 1978 figure 16.1 percent.

The increasing proportion of fringe benefits in total compensation has been attributed to a variety of factors. Robert Rice, in a pioneering effort to explain the growth of wage supplements, listed the following factors: 1) preferential treatment under federal personal income tax laws (an explanation pursued by Martin Feldstein and Donald Cymrot); 2) savings that are made possible by group purchase of some benefits, notably insurance; 3) efforts to reduce turnover in the face of rising costs of labor turnover; and

4) unionization.<sup>2</sup> To Rice's list may be added at least three other factors: 5) preferential treatment under federal corporate income tax laws;<sup>3</sup> 6) the changing age composition of

<sup>2</sup>Rice is somewhat vague about the mechanism by which unions influence the wage-fringe mix, but the mechanism is not far to seek. First, of course, fringe benefits must be bargainable subjects under the law before unions can have any impact. This issue was settled in 1948 in *Inland Steel v. NLRB* (170 F.2d 247 7th Circuit), which made pension plans a mandatory subject of bargaining in the private sector (see Charles Gregory and Harold Katz, p. 629). But, second, given that unions have the potential to alter the wage-fringe mix, we need a theory about why they might do so in fact. Richard Freeman has relied on a median voter model to explain why unions should skew the wage-fringe mix toward fringes.

<sup>3</sup>Payments by employers to pension funds and insurance benefits are deductible from the employer's gross income for tax purposes. The Revenue Act of 1943 is where this favorable tax treatment was first set out comprehensively, although the history of the deductibility of employer contributions is rather subtle (see Dan McGill, pp. 23–26). In a similar vein, see Charles Clotfelter.

the labor force; and 7) the effect of rising incomes (apart from the consequences rising incomes have for pushing households into higher income tax brackets).<sup>4</sup>

A model designed to yield information about the separate effects of each of the above factors on the wage-fringe mix is developed and implemented below. The discussion centers on two questions. First, how well do fringe benefits serve as substitutes for wage benefits in the worker's eyes? Second, how much of the shift in the wage-fringe mix of recent years can be attributed to rising marginal tax rates, how much to rising incomes, and how much to other influences, such as unionism?

In that it represents an explicit attempt to estimate workers' preferences for wage and nonwage benefits, the approach taken here differs from earlier attempts to measure tradeoffs among various components of compensation. Previous research—notably by Richard Thaler and Sherwin Rosen, Joseph Antos and Rosen, Robert E. B. Lucas, Charles Brown, Ehrenberg, and Bradley Schiller and Randall Weiss, among others—has concentrated on measuring market equilibrium tradeoffs among various components of compensation; that is, what Rosen calls envelope functions or market-clearing implicit price curves. As Rosen notes, "An envelope function by itself reveals nothing about the underlying members that generate it" (p. 44). In contrast, the purpose of this paper is to reveal the preferences of workers for wage and nonwage benefits.

### I. Modelling Worker Preferences for Components of Compensation

Consider that there are only two forms of compensation, wages ( $z_w$ ), and fringe benefits ( $z_f$ ), and that wages are taxed at some marginal rate  $t$ , which is a function of household income. Consider further that the worker's utility is a function of the quantities of wages and fringes received in a given time period,

$$(1) \quad U = U(z_w, z_f),$$

and that  $U$  is a well-behaved utility function.

Dual to the direct utility function  $U$  is an indirect utility function  $V$  in the price of wages ( $p_w$ ), the price of fringes ( $p_f$ ), and total compensation ( $M$ ),

$$(2) \quad V = V(p_w, p_f, M),$$

where  $V$  by definition gives the maximum utility attainable by an individual facing given prices of the components of compensation, and given income. It embodies precisely the same information as the direct utility function because it is dual to it. (See, for example, Louis Phlips and Lawrence Lau).

The notion of a "price" of wages or of fringe benefits may appear unorthodox, but must be implicit in any discussion of a quantity of benefits going to a worker. Consider Figure 1, which depicts a representative worker's preferences for wages and fringes, and that worker's problem of choosing a mix of wages and fringes for his or her compensation.<sup>5</sup> In the initial situation, the employer has offered the worker a maximum of  $z_{w0}$  in wages (where  $z_{w0}$  is a quantity of wages),  $z_{f0}$  in fringe benefits (where  $z_{f0}$  is a quantity of fringes), or any combination of  $z_w$  and  $z_f$  lying on the locus

$$(3) \quad z_w(q_w^0) + z_f(q_f^0) = M,$$

where  $M$  is the total compensation of the worker, and  $(q_w^0/q_f^0)$  is the rate at which this worker's employer is willing to exchange wages for fringe benefits. In general, then,  $q_w$  is the employer's price of wages,  $q_f$  the employer's price of fringes.

But the initial constraint described by (3) is not the constraint finally faced by the worker. Rather, suppose that the worker's wages are taxed at some marginal rate,  $t$  (which is constant in the figure), but that employer-paid fringe benefits are untaxed.<sup>6</sup>

<sup>5</sup>That the worker may be a part of a group engaged in collective bargaining, or that he or she may not have individual discretion over the mix of wages and fringes is ignored here for simplicity. But see Gerald Goldstein and Mark Pauly, who take up these complexities.

<sup>6</sup>In fact, of course, the marginal tax rate is not constant, but varies with income. Thus, the budget constraint facing an individual in a progressive tax system is nonlinear, as has been long recognized in the labor supply literature—see, for example the early treat-

<sup>4</sup>See Richard Lester.

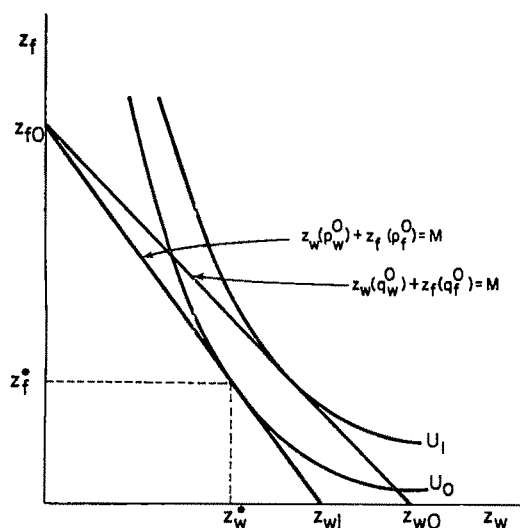


FIGURE 1. THE WORKER'S PROBLEM OF CHOOSING AN OPTIMAL MIX OF WAGES AND FRINGE BENEFITS

The budget constraint pivots to  $z_w(p_w^0) + z_f(p_f^0) = M$ , and the worker optimizes subject to the tax-modified constraint, taking the compensation bundle  $(z_w^*, z_f^*)$ . Here,  $(p_w/p_f)$  is defined as the rate at which the worker is able to trade wages for fringe benefits, or in general,  $p_w$  is the tax-modified price of wages, and  $p_f$  is the price of fringe benefits. Note that the relation between  $(p_w/p_f)$  and  $(q_w/q_f)$  is simply

$$(4) \quad p_w/p_f = (q_w/(1-t))/q_f.$$

For the simple case of  $q_w = q_f = 1$ , we have  $(p_w/p_f) = 1/(1-t)$ .

Now notice (with an eye to estimating an indirect utility function) that in a setting

ments by Terence Wales and Robert Hall. Wales and A. D. Woodland, in the most complete treatment of the nonlinear constraint problem to date, have clearly shown (p. 86) that using a linear approximation to the nonlinear constraint leads in the stochastic case to a specification error and in turn an endogeneity problem—the marginal tax rate observed for each worker is in effect chosen by the worker. In Monte Carlo experiments comparing the linear approximation method with an improved full-information approach they devise, Wales and Woodland find that the linear approximation method tends to underestimate true elasticities of substitution. This result strengthens my findings below of high elasticities of substitution, results which come from the linear approximation method.

where the marginal tax rate varies as a function of income, we will observe variation in the price ratio  $(p_w/p_f)$  facing different individuals or groups in any sample where there is variation in taxable income. Since the marginal tax rate function is a discontinuous or step function of income, it is further possible to consider changes in income apart from changes in the relative price of wages and fringes (since income and the price ratio are not perfectly collinear).

The above discussion is intended to make plausible the notion of a worker's utility function or structure of preferences for wage and nonwage benefits. A more complete specification would of course include leisure (or hours of work) as an argument in the utility function, but to omit leisure simply amounts to assuming separability of leisure and the components of compensation in the worker's utility.<sup>7</sup> The assumption may or may not be onerous depending on the problem and the data at hand. As long as we focus on full-time workers, the assumption is quite harmless.

## II. Estimation

Estimation of the structure of the general direct or indirect utility functions ((1) or (2)) will yield estimates of the structure of workers' preferences for wage and nonwage compensation, and hence yield estimates of numerous behavioral parameters: the elasticity of substitution between wages and fringes, own-price and cross-price elasticities of demand for components of compensation, and income elasticities of demand for components of compensation. These elasticity estimates will embody answers to the questions posed earlier about substitution between wage and nonwage benefits, and the effect of certain variables on the observed mix of wages and nonwage benefits. What is required now is an estimable specification for

<sup>7</sup>See B. K. Atrostic for a study that includes job characteristics and leisure in the worker's utility function. Because Atrostic's concern is with nonpecuniary job characteristics, rather than pecuniary nonwage benefits, she is forced to use principle components analysis and sample partitioning to estimate the price of nonwage benefits, rather than the method of measuring a tax-price of wages used here.

the indirect utility function (2) and data, and these are discussed in turn.

### A. Model Specification

In general, an indirect utility function in the price of wages ( $p_w$ ), the price of fringes ( $p_f$ ), income ( $M$ ), and a vector of control variables ( $x_k$ ) characterizing the worker and his or her workplace may be written as follows:

$$(5) \ln V = \ln V(p_w, p_f, M; x_1, \dots, x_k, \dots, x_K).$$

To estimate the structure of the general indirect utility function, a specification meeting the following criteria must be selected. First, it must be flexible enough to allow a test of homotheticity; that is, we do not want to assume unitary expenditure elasticities. This criterion rules out the Constant Elasticity of Substitution (CES) indirect utility function. Second, it must easily accommodate control variables such as unionization, firm size, and other nonprice variables. This second criterion rules out both the CES and some flexible functional forms. The Generalized Leontief, for example, calls for full interaction of control variables with prices in the estimated demand functions, which is highly unattractive from the point of view of estimation. The Transcendental Logarithmic indirect utility function (see Laurits Christensen, Dale Jorgenson, and Lau) meets both criteria, however, and is hence used.

The translog specification of the arbitrary indirect utility function (5) is

$$\begin{aligned} (6) \quad \ln V = & b_0 + b_f (\ln p_f^*) \\ & + b_w (\ln p_w^*) + \sum_{k=1}^K c_k x_k \\ & + \frac{1}{2} d_{ff} (\ln p_f^*)^2 + d_{fw} (\ln p_f^*) (\ln p_w^*) \\ & + \sum_{k=1}^K g_{fk} (\ln p_f^*) (x_k) \\ & + \frac{1}{2} d_{ww} (\ln p_w^*)^2 + \sum_{k=1}^K g_{wk} (\ln p_w^*) (x_k) \\ & + \sum_{k=1}^K \sum_{l=1}^K h_{kl} (x_k) (x_l), \end{aligned}$$

where  $b_i$ ,  $c_k$ ,  $d_{ij}$ ,  $g_{ik}$ , and  $h_{kl}$  are parameters characterizing the structure of preferences, and  $p_i^* \equiv (p_i/M)$ . The symmetry condition  $d_{wf} = d_{fw}$  has been imposed as (6) is written. Further constraints always imposed are  $g_{fk} = -g_{wk}$  for all  $k$ , and  $h_{kl} = h_{lk}$  for all  $k$  and  $l$ .

The logarithmic form of Roy's identity,

$$(7) \quad S_i = - \frac{\partial \ln V}{\partial \ln p_i} / \frac{\partial \ln V}{\partial \ln M}, \quad i = w, f,$$

(where  $S_i$  is the proportion of compensation taken in form  $i$ ), may be applied to (6) to obtain the following system of demand equations:

$$\begin{aligned} (8) \quad S_i = & \left[ b_i + d_{ii} (\ln p_i^*) + d_{ij} (\ln p_j^*) \right. \\ & \left. + \sum_{k=1}^K g_{ik} (x_k) \right] / D, \end{aligned}$$

for  $i = w, f$ , where<sup>8</sup>

$$\begin{aligned} (9) \quad D = & (b_w + b_f) + (d_{ff} + d_{fw}) (\ln p_f^*) \\ & + (d_{ww} + d_{fw}) (\ln p_w^*). \end{aligned}$$

Since we can observe all the variables in the compensation share equations (8), those equations can be estimated, and we may infer from them the structure of preferences for wage and nonwage benefits.

In addition to exhibiting symmetry of cross-substitution effects (which we have already imposed by setting  $d_{wf} = d_{fw}$ ), the demand system (8) should be homogeneous of degree zero in prices and income, and should satisfy the "adding-up" restriction of consumer theory ( $\sum_i S_i = 1$ ). These restrictions may be considered for, first, a general case and, second, the special case where preferences are homothetic.

<sup>8</sup>The constraint  $g_{fk} = -g_{wk}$  causes the nonprice variables to drop out of  $D$ . The constraint in effect forces the impact of any such characteristic on the share of compensation taken as fringes to be equal and opposite the impact of that characteristic on the share of compensation taken as wages.

For the general case, homogeneity is imposed by setting  $\sum_i \sum_j d_{ij} = 0$ ; that is,  $d_{ww} + d_{ff} = -2d_{fw}$ . Adding-up is obtained by imposing the normalization  $b_w + b_f = -1$ .<sup>9</sup>

For the homothetic case, the condition  $\sum_j d_{ij} = 0$  is imposed for all  $i$ . For our case, this implies that  $d_{ff} = -d_{fw}$  and  $d_{ww} = -d_{fw}$ , so only one independent substitution parameter remains. Adding-up may be obtained by imposing the normalization  $b_w + b_f = 1$ .<sup>10</sup> These three restrictions together imply that  $D=1$ , so that the share equations (8) become linear in parameters.

A further implication of imposing the general restrictions of consumer theory is that the two share equations (8) are not independent. For the general case, either may be estimated by an interactive nonlinear technique. For the homothetic case, either may be estimated by ordinary least squares. Results will be identical regardless of which equation is estimated.<sup>11</sup>

### B. Data

The indirect utility function is estimated below using two quite different sets of data, one of subaggregate groups from surveys published by the BLS, the other of school districts surveyed in the *Census of Governments*.

The BLS has published biennially since 1966 its survey of *Employee Compensation in the Private Nonfarm Economy*, which includes data on the compensation and degree of unionization of four employee groups (office employees in manufacturing industries, office employees in nonmanufacturing industries, nonoffice employees in manufacturing industries, and nonoffice employees in nonmanufacturing industries) classified by

three establishment sizes (firms with fewer than 100, 100–499, and more than 499 employees). Thus, for each of twelve employee groups in each of five years (1966 to 1974), we know average hourly total compensation, the proportions of various components in total compensation, the average number of hours worked per year, and the proportion of employees covered by a collective bargaining contract.<sup>12</sup>

From these data, the variables needed to estimate equation (8) are obtained as follows.

1) The marginal tax on wage income and the tax-price of wages are calculated in the following way: Yearly gross earnings for the average member of each group is computed, and taxable income is figured by taking three exemptions and a standard deduction (or a low income allowance if it is greater).<sup>13</sup> The marginal tax for each group is figured by applying the federal individual income tax rate schedule to the taxable income figures obtained.<sup>14</sup> Finally, the tax-price of wages is

<sup>12</sup>Data from the 1976 BLS *Employee Compensation* survey are excluded from the pooled sample for two reasons. First, unlike earlier surveys, the 1976 survey excluded establishments with fewer than 20 employees, posing a problem of comparability with the earlier surveys. Second, ending the sample in 1974 avoids the problem of accounting for the effects of ERISA, which was signed into law in 1974.

<sup>13</sup>See U.S. Department of Treasury, Internal Revenue Service for the tax regime in each year. Inspection of the tabulations in these documents indicates that using three exemptions and the minimum standard deduction is a reasonable way to calculate taxable income for these subaggregate groups, although a more complicated approach could be imagined.

<sup>14</sup>The issues of 1) accounting for multi-earner families and 2) the appropriate marginal tax rate schedule to use appear to be quite important when using aggregate data. To test for the sensitivity of results to different tax schedules, I calculated marginal taxes and tax-prices of wages ( $p_w/p_f$ ) by applying the taxable income figure to both the tax schedule for married persons filing separately (hereafter "Tax Scheme I") and for married persons filing joint returns ("Tax Scheme II"). Use of Tax Scheme II results in very little increase over time in the marginal tax rate calculated for the groups in the *Employee Compensation* data, because a group's average taxable income must pass some threshold before the group's marginal tax changes. Yet clearly, some individuals in any group face higher marginal taxes as their taxable earnings rise, and this should be reflected in the calculated group marginal tax rates. That it is not is an aggregation problem. Thus, Tax Scheme I seems the proper schedule to use on two counts. First, there is

<sup>9</sup>Both of these normalizations are made with the understanding that the  $p_i^*$  are constructed with means equal to one, so that the desired properties hold at the point of approximation.

<sup>10</sup>Letting the  $b_i$  sum to unity (rather than  $-1$ ) for the homothetic case is attractive because the result is a familiar and easily interpreted linear model. I follow others who have estimated the translog indirect utility function (Christensen, Jorgenson, and Lau; Christensen and Marilyn Manser) in setting  $\sum_i b_i = -1$  for the general case.

<sup>11</sup>For a discussion of this invariance problem, see Ernst Berndt and Christensen.

calculated by the simple formula  $(p_w/p_f) = 1/(1-t)$ , where  $t$  is the marginal tax rate. This formula embodies the assumption that the employer's cost of wages relative to fringes ( $q_w/q_f$ ) is unity, an assumption that may be questionable, but one for which I attempt to correct by including firm-size control variables among the  $x_k$ 's when equation (8) is estimated.

2) Total compensation ( $M$ ) is calculated as a by-product of the marginal tax calculation: first, the tax bill for the average member of each group is calculated by applying the federal individual income tax schedule to taxable income. Second, direct after-tax earnings are calculated by subtracting the total tax bill from yearly gross earnings. Third, direct after-tax payments and voluntary contributions by employers to retirement, health insurance, and life insurance plans are summed to obtain total nominal compensation ( $M'$ ). Finally, this last figure is adjusted by the Consumer Price Index to reflect real after-tax total compensation ( $M$ ).<sup>15</sup>

3) Shares of compensation received as wages ( $S_w$ ) and as fringe benefits ( $S_f$ ) are calculated by dividing direct after-tax payments to workers (for the case of  $w$ ) or employer contributions to pension and to health and life insurance plans (for  $f$ ) by the sum of direct after-tax payments and employer contributions to voluntary plans ( $M'$ ). Thus, employer contributions to legally mandated programs are excluded from consideration because they cannot be affected by the decisions of workers and employers.

4) In addition to including the percentage of workers covered by a collective bargaining

evidence that marginal tax rates facing most households rose between 1966 and 1974 (see U.S. Department of Treasury), and this rise is observed when Tax Scheme II is used. And second, an increasing proportion of households have more than one earner, which means that using the marginal tax on one worker's income calculated as if it were the household's only source of income (i.e., using the joint-filing schedule) will understate the rate at which that income is taxed at the margin.

<sup>15</sup>Ordinarily, a nominal measure of income would be used. In this case, however,  $(p_w/p_f)$  is a tax-price that is not a relative commodity price in the usual sense.

TABLE 2—DESCRIPTIVE STATISTICS FOR BLS  
EMPLOYEE COMPENSATION SURVEY SAMPLE

Variable	Mean	Standard Deviation
Wage Share	0.936	0.021
Fringe Share	0.064	0.021
Proportion Unionized	0.258	0.253
Proportion aged 16-34	0.415	0.029
$\ln(p_w/p_f)$	0.260	0.043
$N = 60$		

contract and dummy variables for establishment size among the  $x_k$ 's, the proportion of the employed labor force aged 16-34 in each year is included to attempt to control for the changing age-composition of the work force. This last is the only variable that is not specific to each of the twelve employee groups; it varies only across years, not across groups.

Descriptive statistics of the variables taken from the *Employee Compensation* survey are displayed in Table 2.

The second set of data used is a sample of independent school districts from the 1977 *Census of Governments* (U.S. Bureau of the Census, 1979a, b, c). In addition to offering data on employer contributions to retirement benefits, and to life and health insurance plans, the *Census of Governments* data have two distinct virtues. First, the unit of observation is the employer (in this case the school district), which is the natural unit of observation for examining a decision that is made between a group of workers and their employer. Second, it is easy to locate each government by state, which makes it possible to take account of state income taxation as well as federal taxation in calculating  $(p_w/p_f)$ .<sup>16</sup> School districts were chosen rather than some other level of government (counties or municipalities, for example) because of the comparative homogeneity of employees both within a school district and across school districts.

<sup>16</sup>This is untrue of the establishment data underlying the BLS *Employee Compensation* survey, which would otherwise be a good alternative source of micro data for the purpose at hand.

There are also at least three disadvantages of using the *Census of Governments* data. First, of the 4,851 school district observations in the sample, only 2,540 have information on employer contributions to benefit plans that appear to be consistent with the information on employee coverage by various plans that also appears in the data file. Second, records from two years and three files of *Census of Governments* data had to be matched and merged.<sup>17</sup> Third, teachers might seem a rather inauspicious group to study because most school districts must under various state statutes participate in a state-administered teachers' (or general public employees') retirement system. In some states, local school districts make no contribution at all to the system, and where they do, that contribution is determined not at the local level but at the state level.<sup>18</sup> Nonetheless, variation in the employer's pension contributions, where they are made, can be observed across states, and variation in contributions to health and life insurance benefits will still be observed over school districts within a state.

With this third problem in mind, two separate samples were constructed and used in estimation: one in which fringe benefits are defined as employer contributions to health and life insurance, the other where they include pension contributions in addition to health and life contributions. The health-life-pension (*HLP*) sample (with 1,499 observations) is considerably smaller than the health-life (*HL*) sample (with 2,540) because all states where school districts need not contribute to a retirement plan (as well as

states where data seemed consistently incomplete) were omitted from the former.<sup>19</sup>

For each of the two samples (*HL* and *HLP*), the variables needed to estimate equation (8) were constructed in a manner similar to that described above for the BLS *Employee Compensation* data, but with the following differences. First,  $S_w$  (the wage share of compensation) for the *HL* sample equals direct after-tax payment to workers divided by the sum of direct after-tax payments and employer contributions to health and life insurance. (And  $S_f = 1 - S_w$ .) For the *HLP* sample,  $S_w$  equals direct after-tax payments divided by the sum of direct after-tax payments, contributions to health and life insurance, and employer contributions to pension plans. Second, both state and local income tax systems were taken into account in calculating direct after-tax payments and the marginal tax rate that (on average) faced each group of teachers.<sup>20</sup> Third, the price index used to adjust total compensation to reflect real after-tax total compensation is a regional price index.<sup>21</sup> Fourth, a different set of control variables ( $x_k$ ) was used in the *Census of Governments* estimation. These were: 1) a set of regional dummy variables for the Northeast, West, and South; 2) a dummy variable equal to one if the school district is in an SMSA, zero otherwise; 3) the enrollment of the school district; 4) the

<sup>19</sup>States included in the health-life-pension sample are Alabama, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Minnesota, Montana, Nebraska, New Hampshire, New Jersey, New Mexico, New York, North Dakota, Pennsylvania, Utah, West Virginia, and Wisconsin. On teacher retirement systems, see Howard Zaritsky, Alan Gustman and Martin Segal, Robert Tilove, Alicia Munnell and Ann Connolly, and U.S. Bureau of the Census (1978).

<sup>20</sup>This was accomplished by a conceptually straightforward but computationally involved algorithm that calculated both federal and state taxable incomes, and applied the federal and appropriate state income tax tables. The total marginal tax is the sum of federal and state marginal income taxes. Information on state tax systems was taken from Advisory Commission on Intergovernmental Relations.

<sup>21</sup>The index used was constructed by John Bishop from BLS figures on living standards (see U.S. Department of Labor) with adjustments by Bishop to reflect different tax rates, transportation costs, and quality-constant housing costs.

<sup>17</sup>Certain data are taken from the 1977 Finance file of the *Census of Governments*, which refers to fiscal year 1977. For school districts, this was usually July (or September) 1976 to June (or August) 1977. The Employment and Labor Relations files of the 1977 *Census* refer, however, to October 1977. Thus the 1977 Finance file data were matched with the Employment and Labor Relations files from the 1976 sample of governments surveyed by the Census. Because the 1976 data are a stratified sample of governments, not a census, there are fewer observations in the sample used here than there would be had we used the *Census*.

<sup>18</sup>Variation within state is, of course, possible if some groups of teachers are able to negotiate the assumption of their contribution by the employer.

TABLE 3—DESCRIPTIVE STATISTICS FOR CENSUS OF GOVERNMENTS SCHOOL DISTRICT SAMPLE, FRINGES DEFINED AS EMPLOYER CONTRIBUTIONS TO HEALTH AND LIFE INSURANCE ONLY

Variable	Mean	Standard Deviation
Wage Share	0.966	0.022
Fringe Share	0.034	0.022
Northeast	0.297	0.457
West	0.278	0.269
South	0.147	0.353
SMSA	0.552	0.497
Enrollment	5803.255	18264.770
Short-Term Debt	190.035	8944.758
Fiscal Dependence	0.485	0.192
Contract	0.701	0.458
Bargaining Statute	0.643	0.479
Threat	0.455	0.222
$\ln(p_w/p_f)$	0.241	0.059
N = 2540		

TABLE 4—DESCRIPTIVE STATISTICS FOR CENSUS OF GOVERNMENTS SCHOOL DISTRICT SAMPLE, FRINGES DEFINED AS EMPLOYER CONTRIBUTIONS TO PENSION FUNDS AND TO HEALTH AND LIFE INSURANCE

Variable	Mean	Standard Deviation
Wage Share	0.912	0.064
Fringe Share	0.088	0.064
Northeast	0.438	0.496
West	0.033	0.177
South	0.050	0.218
SMSA	0.554	0.497
Enrollment	5198.83	18130.92
Short-Term Debt	489.81	10048.90
Fiscal Dependence	0.463	0.182
Contract	0.764	0.425
Bargaining Statute	0.776	0.417
Threat	0.501	0.191
$\ln(p_w/p_f)$	0.247	0.063
N = 1506		

change in the short-term debt position of the school district during the 1977 fiscal year; 5) a measure of the school district's fiscal dependence, defined as the proportion of total revenues originating outside the district; 6) a dummy variable equal to one if there was a collective bargaining contract covering teachers in the district; 7) a dummy variable equal to one if the district was in a state with a statute enabling teachers to bargain collec-

tively and a state agency to administer that statute; and 8) the proportion of teachers in the state covered by a collective bargaining contract, that is, a variable intended to pick up threat and spillover effects of bargaining.

Descriptive statistics for the *Census of Governments* school district samples (both HL and HLP) are displayed in Tables 3 and 4.

### III. Results of Estimation

Table 5 displays the results of estimating equation (8) using the BLS *Employee Compensation* data under both the homothetic and nonhomothetic specifications. Further, each specification is estimated with and without a linear time trend, which is in neither case significant.

The qualitative effects of the control variables on the share of benefits received as wages (the dependent variable) are generally as expected. First, larger establishments pay a smaller proportion of their benefits as wages, supporting the hypothesis that large firms are able to provide fringes at a lower cost than are smaller firms (i.e.  $(q_w/q_f)$  is inversely related to firm size). Specifically, the share of total compensation paid as fringe benefits is 3.0 to 3.5 percentage points higher in firms with 500 or more workers than in firms with fewer than 100 workers.<sup>22</sup> Second, in accord with Freeman's hypotheses and findings, collective bargaining coverage significantly shifts the mix of total compensation toward fringes and away from wages. On average, full collective bargaining coverage increases the fringe benefit share of compensation by 1.6 to 2.2 percentage points compared to a nonunion setting other things equal. Third, there is a negative relation between the proportion of workers aged 16–34 in the labor force, and the wage share of total compensation. This is probably because some fringe benefits—notably health and life insurance contributions—are fixed costs to the employer and are untied to workers'

<sup>22</sup> Recall that the nonhomothetic model is nonlinear. Hence, to calculate the relation between an independent variable and  $S_w$ , one must take  $\partial S_w / \partial x_k$ . Since  $D = -1$  at the means of the independent variables,  $\partial S_w / \partial x_k$  will equal the negative of the coefficient on  $x_k$ .

TABLE 5—PARAMETER ESTIMATES FOR THE TRANSLOG INDIRECT UTILITY FUNCTION APPLIED TO BLS EMPLOYEE COMPENSATION DATA, 1966–74<sup>a,b</sup>

Parameter	Homothetic		Nonhomothetic	
$b_w$	1.0651 (0.0144)	1.0513 (0.0796)	-1.0800 (0.0135)	-1.1291 (0.0739)
$b_f$	-0.0651 (-)	-0.0513 (-)	0.0800 (-)	0.1291 (-)
$d_{ww}$	-0.1508 (-)	-0.1508 (-)	-0.0185 (0.0533)	-0.0256 (0.0546)
$d_{ff}$	-0.1508 (-)	-0.1508 (-)	0.0442 (0.0403)	0.0398 (0.0410)
$d_{fw}$	0.1508 (0.0320)	0.1508 (0.0332)	0.0129 (-)	0.0071 (-)
$g_{wa}$	-0.2610 (0.0352)	-0.2278 (0.1917)	0.3010 (0.0333)	0.4194 (0.1784)
$g_{wu}$	-0.0157 (0.0051)	-0.0157 (0.0052)	0.0216 (0.0049)	0.0219 (0.0049)
$g_{wm}$	-0.0153 (0.0020)	-0.0153 (0.0021)	0.0123 (0.0020)	0.0121 (0.0020)
$g_{wl}$	-0.0343 (0.0025)	-0.0343 (0.0026)	0.0299 (0.0026)	0.0297 (0.0026)
$g_{wt}$	0 (-)	-0.0007 (0.0039)	0 (-)	-0.0024 (0.0036)
$RSS$	0.0021	0.0021	0.0017	0.0016
$R^2$	0.919	0.919	0.936	0.936
$F$	122.15	99.97	157.77	130.25
$N = 60$				

<sup>a</sup>Tax Scheme I was used to compute marginal tax rates used in this estimation. Subscripts refer to wages ( $w$ ), supplements or fringe benefits ( $f$ ), percentage of employed workers aged 16 to 34 ( $a$ ), percent of workers covered by a collective bargaining contract ( $u$ ), medium-sized establishment ( $m$ ), large establishment ( $l$ ), and time trend ( $t$ ). Note that  $g_{fi} = -g_{wi}$  for all  $i$ . The dependent variable is  $S_w$ , the share of compensation received as wages.

<sup>b</sup>Asymptotic standard errors are shown in parentheses. Parameters calculated from restrictions placed on the model are indicated by (-).

wages. They are thus a higher proportion of the salary of a young, low-wage worker than of an older, high-wage worker.

Homotheticity is rejected strongly using an  $F$ -test that compares the homothetic and nonhomothetic specifications.<sup>23</sup> This would seem to support the argument, adumbrated by Lester, that rising incomes increase the value of time to workers, and that higher-income workers hence prefer others to seek out and administer insurance and pension systems for them.

Table 6 displays own-price elasticities ( $\eta_{ff}$  and  $\eta_{ww}$ ), cross-price elasticities ( $\eta_{fw}$  and

$\eta_{wf}$ ), income elasticities ( $\eta_{fM}$  and  $\eta_{wM}$ ), and the substitution elasticity ( $\sigma_{wf}$ ) calculated from the coefficients shown in Table 5. (The estimates actually used are from the equation without the time trend.) Formulas used to compute the figures in Table 6 are given by Christensen and Manser.

Note that two kinds of mean elasticity are shown. The first is the elasticity (for example, the elasticity of substitution,  $\sigma_{wf}$ ) calculated at the sample mean of the independent variables and using sample mean shares. The associated standard error is calculated by taking a second-order Taylor-series expansion about that mean (that is, about the point of approximation). The second set of mean elasticities shown is simply calculated by computing the elasticity (again, suppose

<sup>23</sup>The critical  $F$  for 1 and 54 degrees of freedom at the 1 percent level of significance is 7.20. The  $F$ -test for the null hypothesis of homotheticity is 14.20.

TABLE 6—PRICE, CROSS-PRICE, INCOME, AND SUBSTITUTION ELASTICITIES CALCULATED FROM ESTIMATED INDIRECT UTILITY FUNCTION APPLIED TO BLS EMPLOYEE COMPENSATION DATA

	Homotheticity Imposed				General Form			
	Mean <sup>a</sup>	Mean <sup>b</sup>	Min	Max	Mean <sup>a</sup>	Mean <sup>b</sup>	Min	Max
$\eta_{ff}$	-3.366 (0.034)	-3.637 (0.928)	-6.086	-2.390	-1.663 (0.638)	-1.750 (0.306)	-2.914	-1.382
$\eta_{ww}$	-1.161 (0.034)	-1.161 (0.004)	-1.169	-1.155	-1.012 (0.050)	-1.012 (0.0004)	-1.012	-1.011
$\eta_{fw}$	2.366 (0.501)	2.637 (0.928)	1.390	5.086	0.170 (0.736)	0.196 (0.089)	0.089	0.534
$\eta_{wf}$	0.161 (0.034)	0.161 (0.003)	0.155	0.169	0.045 (0.043)	0.045 (0.0003)	0.045	0.046
$\eta_{fM}$	1.000 (-)	1.000 (-)	1.000	1.000	1.492 (0.131)	1.554 (0.217)	1.294	2.380
$\eta_{wM}$	1.000 (-)	1.000 (-)	1.000	1.000	0.966 (0.009)	0.966 (0.001)	0.965	0.968
$\sigma_{wf}$	3.528 (0.535)	3.798 (0.925)	2.559	6.242	1.674 (0.688)	1.762 (0.307)	1.393	2.926

<sup>a</sup>Elasticity at the mean of the independent variables, and the standard error of the elasticity at that point (in parentheses).

<sup>b</sup>Mean of the elasticities calculated for all observations in the sample, and the associated standard errors are shown below in parentheses.

$\sigma_{wf}$ ) for each observation in the sample, and then taking the mean. The minima and maxima of these computed elasticities are also shown.

The elasticities all have the proper signs and are of reasonable magnitude. Note that the elasticity of substitution ( $\sigma_{wf}$ ) is only half as large in the general nonhomothetic case as with homotheticity imposed. The rejection of the hypothesis of homotheticity implies that, with homotheticity maintained, we wrongly attribute to changes in the tax-price of wages ( $p_w/p_f$ ) changes in the mix of total compensation that are in fact due to changes in income. Thus, the income elasticity of demand for fringes ( $\eta_{fM}$ ) is significantly greater than one in the general case. Even for the general case, though,  $\sigma_{wf}$  is greater than unity (although insignificantly), implying easy substitution of wages for fringe benefits.

Tables 7 and 8 display the results of estimating equation (8) using the *Census of Governments* data, considering life and health insurance alone as fringes (Table 7), and pensions plus life and health insurance as fringes (Table 8).

Looking first at the control variables in the health-life sample, note that being in the

Northeast, being in an urban area, having a collective bargaining contract, and being in a state where there is a heavy concentration of teacher bargaining are all significantly related to a higher proportion of compensation received as fringe benefits. The results are quite different for the health-life-pension sample, where again location in the Northeast and an urban area are associated with a higher proportion of fringe benefits, but where the only significant collective bargaining variable is the presence of a bargaining statute. The insignificance of contracts should come as no surprise in view of the limited scope for local collective bargaining to influence teachers' retirement benefits. The significance of the bargaining statute variable is interesting, though, in that it could imply that prowess in lobbying for a bargaining law is correlated with prowess in influencing state legislatures to adopt attractive public employee retirement systems.

Note that both Tax Scheme I (the tax schedule for married persons filing separately) and Tax Scheme II (the schedule for married persons filing joint returns) have been used to estimate equation (8) with the *Census of Governments* data. That the results

TABLE 7—PARAMETER ESTIMATES FOR THE TRANSLOG INDIRECT UTILITY FUNCTION  
APPLIED TO 1977 CENSUS OF GOVERNMENTS SCHOOL DISTRICT DATA;  
FRINGES DEFINED AS EMPLOYER CONTRIBUTIONS TO HEALTH AND LIFE INSURANCE ONLY<sup>a,b</sup>

Parameter	Tax Scheme I		Tax Scheme II	
	Homothetic	Nonhomothetic	Homothetic	Nonhomothetic
$b_w$	0.9785 (0.0021)	-0.9901 (0.0122)	0.9802 (0.0021)	-0.9867 (0.0116)
$b_f$	0.0215 (-)	-0.0099 (-)	0.0198 (-)	-0.0133 (-)
$d_{ww}$	-0.0087 (-)	0.0024 (0.0096)	-0.0206 (-)	0.0174 (0.0100)
$d_{ff}$	-0.0087 (-)	0.0076 (0.0071)	-0.0206 (-)	0.0203 (0.0083)
$d_{fw}$	0.0087 (0.0070)	0.0050 (-)	0.0206 (0.0082)	0.0189 (-)
$g_{w-North}$	-0.0023 (0.0010)	0.0022 (0.0010)	-0.0023 (0.0010)	0.0022 (0.0010)
$g_{w-West}$	0.0001 (0.0015)	0.0002 (0.0015)	0.0001 (0.0015)	-0.0001 (0.0015)
$g_{w-South}$	0.0182 (0.0015)	-0.0182 (0.0015)	0.0182 (0.0015)	-0.0182 (0.0015)
$g_{w-SMSA}$	-0.0049 (0.0008)	0.0047 (0.0009)	-0.0049 (0.0008)	0.0047 (0.0009)
$g_{w-size}$	0.0000 (0.0000) <sup>b</sup>	-0.0000 (0.0000) <sup>b</sup>	0.0000 (0.0000) <sup>b</sup>	-0.0000 (0.0000) <sup>b</sup>
$g_{w-debt}$	0.0000 (0.0000) <sup>b</sup>	-0.0000 (0.0000) <sup>b</sup>	0.0000 (0.0000) <sup>b</sup>	-0.0000 (0.0000) <sup>b</sup>
$g_{w-dep}$	-0.0015 (0.0022)	0.0025 (0.0022)	-0.0015 (0.0022)	0.0018 (0.0022)
$g_{w-contract}$	-0.0055 (0.0011)	0.0054 (0.0011)	-0.0055 (0.0011)	0.0054 (0.0011)
$g_{w-law}$	-0.0013 (0.0011)	0.0016 (0.0011)	-0.0013 (0.0011)	0.0013 (0.0010)
$g_{w-threat}$	-0.0081 (0.0025)	0.0080 (0.0025)	-0.0081 (0.0025)	0.0080 (0.0025)
RSS	0.9494	0.9490	0.9476	0.9475
R <sup>2</sup>	0.239	0.240	0.241	0.241
F	75.35	66.39	72.91	66.84
N = 2,540				

<sup>a</sup>The dependent variable is  $S_w$ , the share of compensation received as wages. Asymptotic standard are shown below in parentheses.

<sup>b</sup>Coefficient insignificantly different from zero using 95 percent confidence criterion.

of using these different tax schemes are not greatly different can be seen from inspection of Tables 9 and 10, which show own-price, cross-price, income, and substitution elasticities computed from the parameter estimates shown in Tables 7 and 8. (Only the elasticities calculated using the mean of each independent variable and the mean shares are shown.) It is not surprising that the  $\sigma_{wf}$  resulting from calculating  $(p_w/p_f)$  with Tax Scheme II is higher than that resulting from using Tax Scheme I. The tax brackets in the

Tax Scheme II schedule (joint filing) are only one-half as wide as those in the Tax Scheme I schedule (separate filing). Thus, more of the variation in shares is attributed to changes in  $(p_w/p_f)$  under Tax Scheme II, and the  $\sigma_{wf}$  calculated is of course higher.

There is a striking difference, however, between the results obtained when fringes are defined as health and life insurance contributions only (Table 9) and those obtained when fringes are defined as health and life insurance plus pensions (Table 10). The most

TABLE 8—PARAMETER ESTIMATES FOR THE TRANSLOG INDIRECT UTILITY FUNCTION  
APPLIED TO 1977 CENSUS OF GOVERNMENTS SCHOOL DISTRICT DATA;  
FRINGES DEFINED AS EMPLOYER CONTRIBUTIONS TO PENSION FUNDS, AND LIFE AND HEALTH INSURANCE<sup>a</sup>

Parameter	Tax Scheme I		Tax Scheme II	
	Homothetic	Nonhomothetic	Homothetic	Nonhomothetic
$b_w$	1.0557 (0.0081)	-1.5000 (0.0398)	1.0738 (0.0079)	-1.4500 (0.0359)
$b_f$	-0.0557 (-)	0.5000 (-)	-0.0738 (-)	0.4500 (-)
$d_{ww}$	-0.4715 (-)	0.1752 (0.0338)	-0.6547 (-)	0.3813 (0.0360)
$d_{ff}$	-0.4715 (-)	0.3750 (0.0230)	-0.6547 (-)	0.5500 (0.0273)
$d_{fw}$	0.4715 (0.0217)	-0.2751 (-)	0.6547 (0.0260)	-0.4657 (-)
$g_{w \cdot North}$	-0.0492 (0.0034)	0.0378 (0.0034)	-0.0545 (0.0033)	0.0429 (0.0033)
$g_{w \cdot West}$	-0.0079 (0.0075)	0.0114 (0.0072)	-0.0025 (0.0072)	0.0054 (0.0070)
$g_{w \cdot South}$	0.0429 (0.0080)	-0.0399 (0.0077)	0.0400 (0.0077)	-0.0379 (0.0075)
$g_{w \cdot SMSA}$	-0.0016 (0.0032)	0.0011 (0.0072)	-0.0012 (0.0030)	-0.0001 (0.0029)
$g_{w \cdot size}$	0.0000 (0.0000) <sup>b</sup>	-0.0000 (0.0000) <sup>b</sup>	0.0000 (0.0000) <sup>c</sup>	-0.0000 (0.0000) <sup>b</sup>
$g_{w \cdot debt}$	0.0000 (0.0000) <sup>c</sup>	-0.0000 (0.0000) <sup>c</sup>	0.0000 (0.0000) <sup>c</sup>	-0.0000 (0.0000) <sup>c</sup>
$g_{w \cdot dep}$	-0.0397 (0.0078)	0.0577 (0.0076)	-0.0258 (0.0075)	0.0450 (0.0074)
$g_{w \cdot contract}$	-0.0035 (0.0036)	-0.0049 (0.0035)	-0.0046 (0.0035)	-0.0030 (0.0034)
$g_{w \cdot law}$	0.0216 (0.0043)	-0.0096 (0.0042)	0.0281 (0.0041)	-0.0172 (0.0041)
$g_{w \cdot threat}$	-0.0060 (0.0085)	-0.0099 (0.0083)	-0.0141 (0.0082)	-0.0000 (0.0080)
$RSS$	3.5925	3.3196	3.3196	3.0914
$R^2$	0.416	0.460	0.460	0.497
$F$	96.58	105.98	115.69	122.98
$N = 1,506$				

<sup>a</sup>The dependent variable is  $S_w$ , the share of compensation received as wages. Asymptotic standard errors are shown in parentheses.

<sup>b</sup>Coefficient significantly different from zero using 95% confidence criterion.

<sup>c</sup>Coefficient insignificantly different from zero using 95% confidence criterion.

obvious difference is in the estimated  $\sigma_{wf}$ 's, which suggest that retirement income plans are a far better substitute for current wages than are health and life insurance benefits. Yet even for the case where fringes are defined solely as health and life insurance, the point estimates of  $\sigma_{wf}$  exceed unity, indicating that such benefits are a good substitute for wages.

A hardly less obvious difference between the *HL* and *HLP* results is that we cannot reject homotheticity for the *HL* sample (un-

der either tax scheme), whereas homotheticity is strongly rejected for the *HLP* sample.<sup>24</sup>

<sup>24</sup>For the *HL* sample, we use the critical  $F$  for 1 and 2527 degrees of freedom at the 5 percent level of significance, which is approximately 3.85. The  $F$ -test for the null hypothesis of homotheticity is 0.914 using the Tax Scheme I estimates, and 0.32 for the Tax Scheme II estimates. For the *HLP* sample, we use the critical  $F$  for 1 and 1493 degrees of freedom at the 5 percent significance level, which is about 3.87. The  $F$ -test for the null of homotheticity is 122.74 using the Tax Scheme I estimates, and 110.21 using the Tax Scheme II estimates.

TABLE 9—PRICE, CROSS-PRICE, INCOME, AND SUBSTITUTION ELASTICITIES  
CALCULATED FROM ESTIMATED INDIRECT UTILITY FUNCTION APPLIED  
TO CENSUS OF GOVERNMENTS SCHOOL DISTRICT DATA; FRINGES DEFINED  
AS EMPLOYER CONTRIBUTIONS TO HEALTH AND LIFE INSURANCE ONLY<sup>a</sup>

	Tax Scheme I		Tax Scheme II	
	Homotheticity Imposed	General Form	Homotheticity Imposed	General Form
$\eta_{ff}$	-1.258	-1.221	-1.610	-1.598
$\eta_{ww}$	-1.009	-1.005	-1.021	-1.019
$\eta_{fw}$	0.258	0.145	0.610	0.556
$\eta_{wf}$	0.009	0.008	0.021	0.021
$\eta_{fm}$	1.000	1.076	1.000	1.042
$\eta_{wM}$	1.000	0.997	1.000	0.999
$\sigma_{wf}$	1.267	1.227	1.632	1.618

<sup>a</sup>Elasticity at the mean of the independent variables. Standard errors have not been calculated.

TABLE 10—PRICE, CROSS-PRICE, INCOME, AND SUBSTITUTION ELASTICITIES  
CALCULATED FROM ESTIMATED INDIRECT UTILITY FUNCTION APPLIED  
TO CENSUS OF GOVERNMENTS SCHOOL DISTRICT DATA; FRINGES DEFINED AS  
EMPLOYER CONTRIBUTIONS TO PENSION FUNDS, AND HEALTH AND LIFE INSURANCE<sup>a</sup>

	Tax Scheme I		Tax Scheme II	
	Homotheticity Imposed	General Form	Homotheticity Imposed	General Form
$\eta_{ff}$	-6.517	-5.185	-8.660	-7.239
$\eta_{ww}$	-1.516	-1.284	-1.716	-1.492
$\eta_{fw}$	5.517	3.044	7.660	5.270
$\eta_{wf}$	0.516	0.391	0.716	0.583
$\eta_{fM}$	1.000	2.141	1.000	1.969
$\eta_{wM}$	1.000	0.893	1.000	0.909
$\sigma_{wf}$	7.032	5.469	9.376	7.732

<sup>a</sup>Elasticity at the mean of the independent variables. Standard errors have not been calculated.

This result weakens the circumstantial evidence found above in support of Lester's suggestion that there is a high income elasticity of demand for insurance benefits. It suggests instead that there is a very high income elasticity of demand for retirement benefits, and that the effect rising incomes have had on the wage-fringe mix has been mainly through substitution of retirement income for current income, not through substitution of current benefits (such as health insurance) for current income.

#### IV. Discussion and Conclusions

The estimates of the elasticity of substitution between wages and fringes,  $\sigma_{wf}$ , in that

they consistently exceed unity, indicate that wages and wage supplements are easily substituted for each other. Using the BLS data,  $\sigma_{wf}$  is estimated to be 3.5 under the homothetic specification, and 1.7 in the more general nonhomothetic case. Estimates of  $\sigma_{wf}$  obtained using the *Census of Governments* data show that when fringes are defined as health benefits and life insurance plus pensions, wages and fringes are extremely good substitutes, with an estimated elasticity of substitution of 7.7. In contrast, when fringes are defined as only health benefits and life insurance, the estimated elasticity of substitution is 1.6. This estimate, though considerably lower, still indicates relative ease of substitution between wages and fringes.

Knowing the elasticity of substitution between wages and fringes allows us to infer the behavior of the wage-fringe mix in response to changes in the marginal tax-price of wages. That is, because we know  $\sigma_{wf}$  exceeds unity, we would expect a shift in the wage-fringe mix toward fringe benefits in response to rising marginal tax rates facing households. Conversely, the predicted effect of recently passed legislation calling for indexing of the federal income tax is a slowing of the rapid shift toward fringe benefits.

It is logical to ask a further interpretative question. How much of the shift in the wage-fringe mix of recent years can we attribute to rising marginal taxes, how much to rising incomes, and how much to other factors? Consider the case of office employees in large manufacturing establishments. Their wage-share of compensation fell from 0.918 in 1966 to 0.885 in 1974, as their tax-price of wages rose from 1.33 to 1.47, and their real incomes from \$10,223 to \$11,439. Using the estimates from Table 5, one would predict on the basis of the price and income changes alone that the wage-share of this group would have fallen from 0.918 to 0.897. That is, we would have expected more than 63 percent of the observed decrease in the wage share of compensation on the basis of purely economic factors.<sup>25</sup>

Other findings of the study are as follows. First, there appears to be a much higher income elasticity of demand for retirement benefits than for health or life insurance benefits. In fact, we cannot reject the hypothesis that the income elasticity of demand for health and life insurance benefits is unity. However, because retirement benefits are such a large proportion of fringe benefits, the income elasticity of demand for all fringes taken together is found to exceed unity.

Second, larger establishments pay a larger proportion of their benefits as supplemental

benefits. This finding reflects the ability of large firms to take advantage of group purchases of some benefits, and to provide fringes at a lower cost than smaller firms.

Third, collective bargaining coverage significantly shifts the mix of total compensation toward supplemental benefits, as we would expect on the basis of Freeman's theory.

The weaknesses of the relatively simple approach taken above are three. First, the problem of measuring the tax price of wages is especially acute when aggregate data are used, but is not obviated by the use of micro data because of the difficulty of accounting for households with more than one earner. Second, the assumption has been implicit throughout that employees value wage supplements at their cost to the employer, which amounts to assuming that groups of workers may effectively negotiate a package of wages and fringes they prefer. Evidence recently published by Ehrenberg (1980), although specific to pension underfunding of unionized municipal workers, suggests that this assumption may not pose a major obstacle.<sup>26</sup> Third, the approach cannot be extended to measuring tradeoffs between various pairs of nonwage benefits (health insurance and pensions, for example), because there is no tax-price wedge to divide them.<sup>27</sup>

Yet despite these limitations, there is nothing to prevent extension of the technique to the study of policy problems like measuring the effects of the Employee Retirement Income Security Act on compensation and its components, and, with appropriate modifications, it may be a suitable tool for such research.

<sup>26</sup>See also Goldstein and Pauly, who treat supplemental benefits as public goods, and develop a variety of models to explain the quantities of benefits chosen by groups of workers.

<sup>27</sup>Note also that limitations of the data inhibit any attempt to control for the effect of firm-specific human capital on the mix of compensation (a control that would be less consequential with the Census data on teachers than with the BLS data). Data limitations also inhibit more than an imperfect attempt to control for the effect of age on the mix of compensation.

<sup>25</sup>How much of the shift we attribute to rising marginal taxes, and how much to rising incomes, would depend on how much credence we place in the strong rejection of homotheticity we found using these aggregate data. See fn. 9 above for a discussion of the aggregation problem in the present context.

## REFERENCES

- Antos, Joseph R. and Rosen, Sherwin, "Discrimination in the Market for Public School Teachers," *Journal of Econometrics*, May 1975, 3, 123-50.
- Atrostic, B. K., "The Demand for Leisure and Nonpecuniary Job Characteristics," *American Economic Review*, June 1982, 72, 428-40.
- Berndt, Ernst R. and Christensen, Laurits R., "The Translog Function and the Substitution of Equipment, Structures, and Labor in U.S. Manufacturing 1929-68," *Journal of Econometrics*, March 1973, 1, 81-114.
- \_\_\_\_\_, Darrough, M. N., and Diewert, W. E., "Flexible Functional Forms and Expenditure Distributions: An Application to Canadian Consumer Demand Functions," *International Economic Review*, October 1977, 18, 651-75.
- Brown, Charles, "Equalizing Differences in the Labor Market," *Quarterly Journal of Economics*, February 1980, 94, 113-34.
- Caves, Douglas W. and Christensen, Laurits R., "Global Properties of Flexible Functional Forms," *American Economic Review*, June 1980, 70, 422-32.
- Chen, Yung-Ping, "The Growth of Fringe Benefits: Implications for Social Security," *Monthly Labor Review*, November 1981, 104, 3-10.
- Christensen, Laurits R., Jorgenson, Dale W., and Lau, Lawrence J., "Transcendental Logarithmic Utility Functions," *American Economic Review*, June 1975, 65, 367-83.
- \_\_\_\_\_, and Manser, Marilyn E., "Cost-of-Living Indexes and Price Indexes for U.S. Meat and Produce, 1947-1971," in Nestor Terleckyj, ed., *Household Production and Consumption*, New York: National Bureau of Economic Research, 1976, 399-446.
- \_\_\_\_\_, and \_\_\_\_\_, "Estimating U.S. Consumer Preferences for Meat With a Flexible Utility Function," *Journal of Econometrics*, January 1977, 5, 37-57.
- Clotfelter, Charles T., "Business Perks and Tax-Induced Distortions: The Case of Travel and Entertainment," Department of Economics Working Paper, Duke University, 1981.
- Cymrot, Donald, J., "Private Pension Saving: The Effect of Tax Incentives on the Rate of Return," *Southern Economic Journal*, July 1980, 47, 179-90.
- Ehrenberg, Ronald J., "Retirement System Characteristics and Compensating Differentials in the Public Sector," *Industrial and Labor Relations Review*, July 1980, 33, 470-83.
- Feldstein, Martin, "The Welfare Loss of Excess Health Insurance," *Journal of Political Economy*, March/April 1973, 81, 251-80.
- Freeman, Richard B., "The Effect of Unionism on Fringe Benefits," *Industrial and Labor Relations Review*, July 1981, 34, 489-509.
- Goldstein, Gerald, S. and Pauly, Mark V., "Group Health Insurance as a Local Public Good," in Richard N. Rosett, ed., *The Role of Health Insurance in the Health Services Sector*, New York: National Bureau of Economic Research, 1976, 73-110.
- Gregory, Charles O. and Katz, Harold A., *Labor and the Law*, 3d ed., New York: Norton, 1979.
- Gustman, Alan L. and Segal, Martin, "Interstate Variations in Teachers' Pensions," *Industrial Relations*, October 1977, 16, 335-44.
- Hall, Robert E., "Wages, Income, and Hours of Work in the U.S.," in Glen G. Cain and Harold W. Watts, eds., *Income Maintenance and Labor Supply*, New York: Academic Press, 1973.
- Lau, Lawrence J., "Duality and the Structure of Utility Functions," *Journal of Economic Theory*, December 1969, 1, 374-96.
- Lester, Richard, A., "Benefits as a Preferred Form of Compensation," *Southern Economic Journal*, April 1967, 33, 488-95.
- Lucas, Robert E. B., "Hedonic Wage Equations and Psychic Wages in the Returns to Schooling," *American Economic Review*, September 1977, 67, 549-58.
- McGill, Dan M., *Fundamentals of Private Pensions*, 4th ed., Homewood: R. D. Irwin and Pension Research Council, 1979.
- Munnell, Alicia H. and Connolly, Ann M., "Funding Government Pensions: State-Local, Civil Service, and Military," in *Funding Pensions: Issues and Implications for Financial Markets*, Federal Reserve

- Bank of Boston Conference Series No. 16, October 1976.
- Phlips, Louis, *Applied Consumption Analysis*, Amsterdam: North-Holland, 1974.
- Rice, Robert G., "Skill, Earnings, and the Growth of Wage Supplements," *American Economic Review Proceedings*, May 1966, 56, 583-93.
- Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, January-February 1974, 82, 34-55.
- Schiller, Bradley R. and Weiss, Randall D., "Pensions and Wages: A Test for Equalizing Differences," *Review of Economics and Statistics*, November 1980, 62, 529-38.
- Smith, Robert S., "Compensating Wage Differentials and Public Policy: A Review," *Industrial and Labor Relations Review*, April 1979, 32, 339-52.
- Thaler, Richard and Rosen, Sherwin, "The Value of Saving a Life: Evidence from the Labor Market," in Nestor E. Terleckyj, ed., *Household Production and Consumption*, New York: National Bureau of Economic Research, 1976, 265-98.
- Tilove, Robert, *Public Employee Pensions*, New York: Columbia University Press, 1976.
- Updike, John, *Rabbit Is Rich*, New York: Alfred A. Knopf, 1981.
- Wales, Terrence J., "Estimation of a Labor Supply Curve for Self-Employed Business Proprietors," *International Economic Review*, February 1973, 14, 69-80.
- \_\_\_\_\_, and Woodland, A. D., "Labor Supply and Progressive Taxes," *Review of Economic Studies*, June 1979, 46, 83-95.
- Zaritsky, Howard, "A Summary and Analysis of the Pension and Retirement Systems of the 50 States and the District of Columbia," in Committee on Education and Labor, U.S. House of Representatives, *Pension Task Force Report on Public Employee Retirement Systems*, Washington: USGPO, 1978.
- Advisory Commission on Intergovernmental Relations, *Significant Features of Fiscal Federalism*, 1976-77 edition, Washington: USGPO, 1977.
- Chamber of Commerce of the United States, *Employee Benefits*, Washington: Chamber of Commerce, various years.
- U.S. Bureau of the Census, *1977 Census of Governments: Employee-Retirement Systems of State and Local Governments*, Washington: USGPO, 1978.
- \_\_\_\_\_, (1979a) *1977 Census of Governments: Compendium of Public Employment*, Washington: USGPO, 1979.
- \_\_\_\_\_, (1979b) *1977 Census of Governments: Finances of School Districts*, Washington: USGPO, 1979.
- \_\_\_\_\_, (1979c) *1977 Census of Governments: Labor-Management Relations in State and Local Governments*, Washington: USGPO, 1979.
- U.S. Department of Labor, Bureau of Labor Statistics, *Employee Compensation in the Private Nonfarm Economy*, Bulletins 1627, 1722, 1770, 1873, 1963, Washington: USGPO.
- \_\_\_\_\_, *Three Standards of Living for an Urban Family of Four Persons*, Bulletin 1570-5, Washington: USGPO.
- U.S. Department of Treasury, Internal Revenue Service, *Statistics of Income: The Personal Income Tax*, Washington: USGPO, various years.

# Exchange Risk, Political Risk, and Macroeconomic Equilibrium

By JONATHAN EATON AND STEPHEN J. TURNOVSKY\*

With the advent of increased exchange rate flexibility, the role of policy and the transmission of economic disturbances under flexible rates have become topics of considerable interest. J. Marcus Fleming (1962) and Robert Mundell (1963), in their early analysis of the theory of policy in open economies, emphasized the importance of the degree of capital mobility. Subsequently, a number of authors have analyzed the implication of the degree of capital mobility for exchange rate and income determination in open economies; see, for example, Pentti Kouri and Michael Porter (1974), William Branson (1975, 1979), Lance Gorton and Dale Henderson (1976), and Turnovsky (1976). The polar case of perfect capital mobility has received most attention.

Capital is perfectly mobile when investors consider domestic bonds, denominated in domestic currency, and foreign bonds, denominated in foreign currency, to be perfect substitutes. Two factors may cause perfect substitutability to break down. One is *exchange risk*: the values of the two bonds are defined in terms of different currencies and the exchange rate at the date of maturation is uncertain. The second is *default risk*: investors may perceive that foreign bonds are more subject to the risk of default than are domestic bonds. For one thing, in the event of default the investor must pursue his claim through a foreign legal system, which is likely to be costly. Robert Aliber (1973) terms differences in default risk arising from the "foreignness" of assets as "political risk."<sup>1</sup>

\*Yale University and University of Illinois, respectively.

<sup>1</sup>See Peter Kenen (1965) for an early analysis of the forward market incorporating potential default. Political risk may derive from the threat of exchange controls as well as from the threat of outright debt repudiation. In either case, the risk may not be the loss of the total net worth of the investment. When the threat derives from exchange controls, the expected value of net worth may even be preserved, but if payment is delayed, forward cover obtained for the investment will no longer be

Where domestic and foreign securities are assumed to be only imperfect substitutes, it is important to distinguish the degree of imperfection due to each of these sources of risk. While this distinction has been noted in optimal portfolio models developed by authors such as Kouri (1976), and Michael Alder and Bernard Dumas (1977), it has not been brought out in the current macro literature. The objective of the present paper is to investigate the macroeconomic consequences of this distinction and to show how it may be crucial in assessing the effects of various macroeconomic disturbances under varying degrees of capital mobility. We find that well-known results about the effects of increased capital mobility on the efficacy of monetary policy are correct when the increased mobility is due to a reduction in exchange risk, but not necessarily so when it is due to a reduction in default risk.

Section I formulates a general equilibrium macroeconomic model which embodies the distinction between these two sources of risk. We assume that the economy is small, so that it treats certain characteristics of the foreign economy as parametric. Section II carries out certain comparative static exercises with this model. Specifically, it considers some important relationships between the available policy instruments and endogenous variables, and shows how these are affected by the degree of capital mobility in the two senses in which this term is defined. In particular, we show that, as long as there is a possibility of default, there exists an independent role for forward market intervention. In the limiting case of no default risk,

appropriate. For our purposes it is the risk that repayment not occur on the date for which the forward cover is arranged, thereby exposing the investor to exchange risk, that is relevant. Our analysis could easily be modified to incorporate a preservation of some or all net worth in the event of default. See Eaton and Mark Gersovitz (1981) for a model in which the probability of default is endogenous.

when domestic and foreign bonds are perfect substitutes on a covered basis, this independent role ceases and forward market intervention becomes equivalent to bond market intervention. Secondly, the comparative statics of changes in the degree of capital mobility are discussed. Apart from the results on the effectiveness of monetary policy noted above, we also consider how varying degrees of capital mobility affect the response of the domestic economy to foreign shocks.

### I. A General Equilibrium Macroeconomic Model

Consider the following macroeconomic model of a small, open economy:

$$(1) \quad b_t - p_t = -\omega_1(i_t^* + e_t^f - e_t^s - i_t) + \omega_2^D i_t + \omega_3^D y_t, \quad \omega_1 > 0, \omega_2^D > 0, \omega_3^D \geq 0,$$

$$(2) \quad \Pi[\omega_1(i_t^* + e_t^f - e_t^s - i_t) + \omega_2^F i_t + \omega_3^F y_t] = \gamma(e_{t+1}^s - e_t^f) + \lambda(g_t - p_t),$$

$$\omega_2^F > 0, \omega_3^F > 0, \gamma > 0,$$

$$(3) \quad m_t - p_t = \alpha_1 y_t - \alpha_2 i_t, \quad \alpha_1 > 0, \alpha_2 > 0,$$

$$(4) \quad p_t = \delta(p_t^* + e_t^s) + (1 - \delta)p_{t-1},$$

$$0 \leq \delta \leq 1,$$

$$(5) \quad y_t = \theta(p_t - p_{t-1}), \quad \theta \geq 0,$$

where  $b$  = domestic nominal stock of bonds,  $m$  = domestic nominal stock of money,  $y$  = domestic real output,  $p$  = domestic price level,  $p^*$  = foreign price level,  $i$  = domestic nominal interest rate,  $i^*$  = foreign nominal interest rate,  $e^s$  = spot price of foreign currency in terms of domestic currency,  $e^f$  = one-period forward price of foreign currency in terms of domestic currency,  $g$  = net government purchases of foreign currency in the forward market, translated to domestic currency units at the forward exchange rate,  $x_t$  = value of variable  $x$  at time  $t$ , and  $x_{t+i}$  = forecast of variable  $x_{t+i}$  formed at time  $t$  ( $x = e^s, p$ ). All variables except  $i$  and  $i^*$  are measured in logarithms.

The specification of these equations and the restrictions embodied in them can be derived as a *log-linear* approximation to an underlying two-period, two-country mean-variance optimizing model.<sup>2</sup> Here we consider the limiting case of this model in which one country, the one we analyze, is small in that it cannot affect  $p_t^*$  and  $i_t^*$ , the foreign price level and interest rate. A brief description of these relationships follows.

Equation (1) describes equilibrium in the domestic bond market for the case in which foreigners hold zero stocks of domestic bonds.<sup>3</sup> In the presence of default risk on foreign bonds, domestic and foreign bonds are gross but imperfect substitutes, so that the demand for domestic bonds varies positively with the domestic interest rate and inversely with the covered foreign rate. It is convenient to introduce these in the equivalent form  $i_t$  and  $(i_t^* + e_t^f - e_t^s - i_t)$ ; the second measuring the deviation from covered interest parity (*CIP*). The coefficient  $\omega_1$  parametrizes the degree of substitutability between domestic and foreign assets stemming from default risk. As default risk tends to zero, the two bonds become perfect substitutes and (1) yields the *CIP* condition  $i_t = i_t^* + e_t^f - e_t^s$ . The income coefficient in equation (1) is ambiguous in sign. An increase in income raises both the demand for money for transactions and the demand for total wealth. The first acts to reduce the demand for domestic bonds while the second acts to increase it.

Forward market equilibrium is described by equation (2). Private individuals participate in the forward market either to ob-

<sup>2</sup>The micro theory underlying the specification of the macro model summarized in equations (1)–(5) is developed in a longer version of this paper (1981), bearing the same title. There we reference existing literature dealing with financial models of currency speculation and foreign investment.

<sup>3</sup>Our portfolio model (1981) indicates that this case is most likely to arise when the domestic price level is more sensitive to the exchange rate than is the foreign price level.

Note we assume (i) that money and bonds are both outside assets and (ii) that the money and bonds of each country are denominated in the currency of that country. Jeffrey Frankel discusses the relevance of these assumptions for our specification.

tain forward cover on their holdings of bonds denominated in foreign currency, or to speculate on changes in relative currency values.<sup>4</sup> The stock of foreign bonds owned domestically is given by the term in parentheses on the left-hand side of (2). It takes a form analogous to that for the demand for domestic bonds, with the coefficients having the signs indicated. Note that the income coefficient is positive, a fact that can be established from the underlying optimization. The coefficient  $\Pi$  is the fraction of foreign bond holdings that are covered, which equals the probability of no default on these bonds.<sup>5</sup>

Forward market equilibrium requires that the net private supply of foreign exchange forward to cover foreign bonds equal the net private speculative demand for foreign exchange forward plus the net government purchases of foreign exchange forward. Equation (2) states this relationship in real terms. Following the portfolio literature, we assume that the real demand for foreign exchange for speculation is a function of the expected rate of return on the forward purchase of foreign currency as described by the first term on the right-hand side of (2). The parameter  $\gamma$  reflects the degree of substitutability between domestic and foreign bonds due to exchange risk and risk aversion, varying inversely with each. As exchange risk disappears or as risk neutrality is approached,  $\gamma \rightarrow \infty$ , implying that  $e_{t+1,t}^s = e_t^f$ . The expected rates of return on covered and uncovered foreign bonds are then equal. Otherwise, the existence of exchange risk creates a divergence between them. Finally, net real government intervention is described by  $\lambda(g_t - p_t)$ , where  $\lambda$  denotes the average

real stock of forward exchange held by government and enters in the process of the log-linearizing the underlying additive relationship.

Money market equilibrium is described by (3). This specification can be derived from a portfolio maximizing problem in which money provides transactions services related to income.<sup>6</sup> Note that only the nominal interest rate on domestic bonds appears. The reason is that, since domestic money and domestic bonds are denominated in the same currency, and since both are free of default risk from the domestic wealthholder's vantage point, the two assets have the same risk characteristics. Thus while the decision to divide total wealth between foreign bonds and all domestic-currency-denominated assets depends upon the risk and return characteristics of the domestic and foreign bonds, the division of domestic-currency-denominated assets between money and bonds depends upon transactions needs alone. The two decisions are separable, and only the domestic interest rate and transactions variable (income) are relevant for the second.

The price equation (4) assumes that the prices of some tradable commodities are established in auction markets, while the prices of other tradable goods and nontradable commodities are established by contracts in the previous period. The prices of the first group are hence flexible within the period and are determined by the international law of one price. The prices of the second group are fixed within the period, and set at the price level expected to obtain in the relevant period as of the previous period. This assumption parallels the contract theory of wage determination of Jo Anna Gray (1976), Stanley Fischer (1977), and Edmund Phelps and John Taylor (1977). The overall price index is therefore a weighted average of these two prices, where  $\delta$  is the share of the flexible-priced commodities.

Finally, equation (5) assumes that the deviation of output from its natural level, normalized by choice of units to be zero, is determined by the unanticipated component of the domestic price level. Such a specifica-

<sup>4</sup>Since foreign bonds are subject to default risk, while we assume that forward contracts are not, exchange rate speculation takes the form of an open position in the forward market rather than borrowing in one currency to buy bonds denominated in the other (compare Adler and Dumas). If there is no default risk associated with either type of transaction, the two are equivalent.

<sup>5</sup>This proposition is established formally in our 1981 paper. Note that our specifications for the demand for domestic bonds, foreign bonds, and money, embodied in equations (1), (2) and (3), respectively, implicitly define a savings function. See Robert E. Lucas (1975), for example, who follows a similar procedure.

<sup>6</sup>We do so in our 1981 paper.

tion follows from the contract theory of wage determination referred to above.<sup>7</sup>

Equations (1)–(5) constitute a complete macroeconomic model of a small, open economy. The system of equations determines equilibrium values of the five endogenous variables  $p_t$ ,  $i_t$ ,  $e_t^s$ ,  $e_t^f$ , and  $y_t$  in terms of the expectations  $e_{t+1,t}^s$ ,  $p_{t,t-1}$ , and the exogenous variables  $m_t$ ,  $b_t$ ,  $g_t$ ,  $i_t^*$ , and  $p_t^*$  as well as the structural parameters of the system. To close the model we assume that expectations are formed rationally; that is, that price and exchange rate expectations are formed on the basis of the model itself and all available information.

We now use the model to contrast the effects of two different ways in which capital can become more mobile. On the one hand, exchange risk, or aversion to exchange risk, may fall. This is reflected by an increase in the parameter  $\gamma$ . As a consequence, the forward rate becomes more closely tied to the expectation of the future spot rate. On the other hand, default risk or aversion to default risk may fall, raising the parameter  $\omega_1$ .<sup>8</sup> As a consequence, the domestic interest rate will be determined more closely by the covered interest parity condition. We show that the way in which capital mobility is

increased has important implications for the response of the economy to a monetary disturbance.

## II. Comparative Statics

In the model described above, individuals' expectations about the future influence their current behavior. As is well known (compare Taylor), when expectations are formed rationally, such models have nonunique solutions. There is only one solution to this model, however, for which the exchange rate and the price level remain bounded into the infinite future. We focus on this solution to the model.

Our purpose in solving the model is to analyze the response of the economy to changes in the policy variables  $m_t$ ,  $b_t$ , and  $g_t$  and also to changes in the foreign variables  $i_t^*$  and  $p_t^*$ . Disturbances in the exogenous variables may be distinguished by a number of characteristics: (a) their expected permanence or transience, (b) the extent to which they are anticipated or unanticipated, (c) whether they occur immediately or are announced in advance. In principle, we could analyze the effects of many specific types of disturbances. We shall limit ourselves, however, to two types. First, we consider the effect on impact of disturbances that are (a) unannounced, (b) unanticipated, and (c) expected to last only one period. Second, we consider the effects in steady state of changes in the exogenous variables that are perceived as permanent.

To analyze the effect of a temporary, unanticipated disturbance in period  $t$ , we may assume, without loss of generality, that  $m_{\tau,t} = b_{\tau,t} = g_{\tau,t} = p_{\tau,t}^* = i_{\tau,t}^* = 0$ ,  $\tau > t$ , implying that  $e_{t+1,t}^s = 0$ , while  $m_{t,\tau} = b_{t,\tau} = g_{t,\tau} = p_{t,\tau}^* = i_{t,\tau}^* = 0$ ,  $\tau < t$ , which implies that  $p_{t,t-1} = 0$ . The reduced-form expressions for the endogenous variables are easily obtained by solving the five equations of the system with these assumptions. The expressions are messy, however, and we do not report them explicitly.

To analyze the steady-state effects of permanent changes in the exogenous variables we set  $p_{t,t-1} = p_t = \bar{p}$  and  $e_{t+1,t}^s = e_t^s = \bar{e}^s$ , where  $\bar{x}$  denotes the steady value of  $x$ . Again

<sup>7</sup>Equation (4) may be obtained in the following way: Assume that all wages are determined by one-period-ahead contracts so that we may set  $w_t = p_{t,t-1}$ , where  $w_t$  is the logarithm of the wage in period  $t$ . In the contract-price sector, the wage divided by the output price is constant as are output and employment. Thus we may set  $y_t^C = 0$ , where  $y_t^C$  is the logarithm of the output of the contract-price sector. In the flexible-price sector, the wage divided by the output price is inversely proportional to the price level. Thus output and employment expand when prices are high and conversely. Defining  $y_t^A$  as the output of the flexible-price sector, we have

$$y_t^A = \theta'(p_t^A - p_{t,t-1}), \quad \theta' \geq 0,$$

where  $p_t^A$ , the price of the flexible-priced commodities, is given by the international law of one price,  $p_t^A = p_t^* + e_t^s$ . This last equation, together with the specification of  $y_t^A$  and (4), implies  $y_t = \theta(p_t - p_{t,t-1})$ , where  $\theta = \theta'/\delta$ .

<sup>8</sup>Strictly speaking, we should also take account of the fact that a reduction in default risk will increase the fraction of foreign bonds that are covered by a position in the forward market. This simply complicates the comparative statics somewhat and does not change our conclusions in any substantive way.

the expressions are tedious and are not reported.

#### A. Money Market, Bond Market, and Forward Market Intervention

The government can affect the domestic economy by changing the supply of money, the stock of domestic bonds, and its forward market position.<sup>9</sup> Manipulation of each instrument impinges initially on different markets and has differential effects on domestic income, the price level, and the interest rate. The three policy instruments are linearly independent except in some special cases. We summarize the most important relationships between policy instruments and endogenous variables:

1) The effect, on impact, of an *unanticipated, transitory* increase in  $m_t$ ,  $b_t$ , or  $g_t$  is expansionary. As is well known, increasing  $m_t$  or  $b_t$  raises nominal wealth above its desired level at initial commodity prices. Prices that are flexible rise, inducing an expansion in income. An increase in  $g_t$ , government holdings of foreign exchange forward, raises  $e_t^f$  and hence the rate of return on *covered* foreign bonds. Individuals' demand for domestic bonds falls so that a higher price level is required to equate demand to the existing nominal stock. For similar reasons, the *steady-state* effect of *permanent* increases in  $\bar{m}$ ,  $\bar{b}$  and  $\bar{g}$  is to raise  $\bar{\pi}$  and  $\bar{e}^*$ ;  $\bar{y}$  remains fixed at its long-run natural level.

2) A permanent increase in  $\bar{m}$  lowers the domestic and covered foreign interest rates, while a permanent increase in  $\bar{b}$  or  $\bar{g}$  does the opposite;  $\bar{b}$  has relatively more impact on the domestic rate while  $\bar{g}$  has relatively greater influence on the covered foreign rate.

3) In the limit as the probability of default goes to zero,  $\omega_1 \rightarrow \infty$  and covered interest parity obtains. In this case,  $b_t$  and  $g_t$  affect all domestic variables in the same linear proportion,  $b_t + \lambda g_t$ . In other words, bond market and forward market intervention be-

come equivalent in their effects on all endogenous variables; either one is therefore redundant. The same applies to  $\bar{b}$  and  $\bar{g}$ .

4) In the limit as exchange risk goes to zero,  $\gamma \rightarrow \infty$  and forward market intervention ceases to have an effect on any endogenous variable. Any forward market intervention by the monetary authority is exactly offset by the actions of speculators.

5) In the limit as both exchange risk and default risk go to zero, (or else as risk aversion goes to zero), both  $\gamma \rightarrow \infty$  and  $\omega_1 \rightarrow \infty$  and neither bond market nor forward market intervention can have any effect. The familiar proposition that fiscal policy (i.e., a change in the government deficit unaccompanied by a change in the money supply) is ineffective under flexible rates and perfect capital mobility (see Mundell, 1963) is true only if perfect capital mobility is interpreted to mean an absence of both default risk and exchange risk. Otherwise fiscal policy is still effective.

#### B. Capital Mobility, the International Transmission of Disturbances, and the Role of Policy

While a number of studies have emphasized the importance of the degree of capital mobility for both the efficacy of policy and the international transmission of disturbances, no distinction has been made in this literature, to our knowledge, between the two ways in which it can occur.

In fact, an increase in capital mobility in either sense tends to increase the sensitivity of the domestic economy to foreign disturbances. The response of the domestic interest rate to a change in the foreign interest rate, whether it is transitory  $[\partial i_t / \partial i_t^*]$  or permanent  $[\partial \bar{i} / \partial \bar{i}^*]$ , increases with an increase in either  $\omega_1$  or  $\gamma$ . Likewise, the response of the domestic price level to a transitory increase in the foreign price level  $[\partial p_t / \partial p_t^*]$  increases with both  $\omega_1$  and  $\gamma$ , while the steady-state response  $[\partial \bar{p} / \partial \bar{p}^*]$  is always zero.

Increased capital mobility in its two different forms, however, can have very different implications for the effects of domestic policies. For example, a reduction in *exchange*

<sup>9</sup>Elsewhere (1980) we discuss how these various forms of intervention are related by the balance sheet constraint of the combined monetary-fiscal authority.

*risk* or aversion to exchange risk always increases the steady-state responses of the exchange rate and price level to permanent changes in the domestic money supply. This is because as  $\gamma$  increases, the steady-state forward and spot rates move closer together, and the covered foreign exchange rate becomes less flexible, as does the domestic interest rate. As a result, more of the change in the money supply must be reflected in a higher price level and exchange rate and less in a lower interest rate.

By contrast, a reduction in *default risk* has an ambiguous effect on the steady-state relationship between the money supply and price level. An increase in  $\omega_1$  raises or lowers the term  $\partial \bar{p} / \partial \bar{m}$  as

$$\alpha_2 \gamma (\gamma - \Pi \omega_2^F - \lambda \omega_2^D) \geq 0.$$

The reason is that an increase in  $\omega_1$  always ties the domestic interest rate more closely to the *covered* foreign interest rate. When  $\gamma$  is large, the covered and uncovered foreign rates are close, so that a reduction in default risk ties the domestic interest rate more closely to the *uncovered* foreign rate as well, so that it is less flexible. As with a reduction in exchange risk, then, more of a change in money supply must be absorbed by the price level rather than by the interest rate. When  $\gamma$  is small, however, the covered and uncovered foreign rates can diverge, and the increase in  $\bar{m}$  can be accompanied by a fall in  $\bar{i}$  and  $\bar{e}^f - \bar{e}^s$ . If  $\omega_2^F$  is large, then the fall in  $\bar{i}$  raises the demand for *covered* foreign bonds. The consequent increase in the sale of foreign exchange forward to cover these bonds requires, for forward market equilibrium, offsetting speculation;  $\bar{e}^s - \bar{e}^f$  must consequently rise to elicit this speculation. The interest rate on covered foreign bonds,  $\bar{i}^* + \bar{e}^f - \bar{e}^s$  therefore falls, inducing a further fall in the domestic interest rate  $\bar{i}$ . Increased capital mobility in this sense can therefore make the domestic interest rate *more* flexible. In other words, an increase in the degree of capital mobility in the sense of domestic and covered foreign bonds becoming closer substitutes can reduce the rise in the price level and increase the change in the interest rate resulting from a monetary expansion.

### III. Conclusion

Exchange risk and political risk both diminish the mobility of capital. Macroeconomic analyses of open economies have emphasized the importance of the degree of capital mobility, but have not, typically, distinguished between these two impediments to mobility. In fact, they impinge on the economy in different ways. On the one hand, exchange risk creates a divergence between the forward exchange rate and the spot rate expected to prevail when the forward contract matures, in turn creating a divergence between the expected rates of return on covered and uncovered foreign bonds. Exchange risk does not, however, affect the relationship between the *covered* foreign interest rate and the domestic interest rate. Political risk, on the other hand, weakens this second relationship, but does not affect the first.

In the absence of both exchange risk and political risk, that is, if capital is perfectly mobile, changes in the supply of domestic bonds or in the government's forward position have no effect on the domestic economy. Only changes in the money supply have an impact. Introducing exchange risk alone makes bond market and forward market intervention effective, although their effects are not linearly independent. They are essentially equivalent, and one or the other is redundant. Alternatively, introducing default risk alone gives efficacy to bond market intervention, but not to forward market intervention. Finally, if both exchange and default risk are present, then forward market, bond market, and money market intervention all have independent effects on the economy.

An increase in capital mobility due to the diminution of the effects of either type of risk increases the sensitivity of the domestic interest rate to the foreign interest rate and the sensitivity of the domestic price level to the foreign price level. An increase in capital mobility in the form of an increase in the elasticity of speculation increases the steady-state effect of a change in the money supply on the domestic price level. However, an increase in capital mobility that takes the

form of decreasing the divergence between the domestic and foreign covered interest rates has an ambiguous effect on the relationship between money and the price level.

## REFERENCES

- Adler, Michael and Dumas, Bernard, "The Microeconomics of the Firm in an Open Economy," *American Economic Review Proceedings*, February 1977, 67, 180-89.
- Aliber, Robert, "The Interest Rate Parity Theorem: A Reinterpretation," *Journal of Political Economy*, November/December 1973, 81, 1451-59.
- Branson, William H., "Portfolio Equilibrium and Monetary Policy with Foreign and Non-Traded Assets," in E. Claassen and P. Salin, eds., *Recent Issues in International Monetary Economics*, Amsterdam: North-Holland, 1975.
- \_\_\_\_\_, "Exchange Rate Dynamics and Monetary Policy," in A. Lindbeck, ed., *Inflation and Employment in Open Economies*, Amsterdam: North-Holland, 1979.
- Eaton, Jonathan and Gersovitz, Mark, "Debt with Potential Repudiation: Theoretical and Empirical Analysis," *Review of Economic Studies*, April 1981, 48, 289-309.
- \_\_\_\_\_, and Turnovsky, Stephen J., "The Forward Exchange Market, Speculation and Exchange Market Intervention," Working Paper No. 033, Australian National University, November 1980.
- \_\_\_\_\_, and \_\_\_\_\_, "Exchange Risk, Political Risk and Macroeconomic Equilibrium," Discussion Paper No. 388, Economic Growth Center, Yale University, September 1981.
- Fischer, Stanley, "Wage Indexation and Macroeconomic Stability," in K. Brunner and A. Meltzer, eds., *Stabilization of the Domestic and International Economy*, Amsterdam: North-Holland, 1977.
- Fleming, J. Marcus, "Domestic Financial Policies under Fixed and Floating Exchange Rates," *IMF Staff Papers*, November 1962, 9, 369-80.
- Frankel, Jeffrey A., "The Diversifiability of Exchange Risk," *Journal of International Economics*, August 1979, 9, 378-93.
- Girton, Lance and Henderson, Dale W., "Financial Capital Movements and Central Bank Behavior in a Two-Country, Short-Run Portfolio Balance Model," *Journal of Monetary Economics*, January 1976, 2, 33-67.
- Gray, Jo Anna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, 2, 221-35.
- Kenen, Peter B., "Trade, Speculation, Aid and the Forward Exchange Rate," in R. Baldwin et al., eds., *Trade, Growth, and the Balance of Payments*, Chicago: Rand-McNally, 1965.
- Kouri, Pentti J. K., "The Determinants of the Forward Premium," Seminar Paper No. 62, Institute for International Economic Studies, University of Stockholm, August 1976.
- \_\_\_\_\_, and Porter, Michael, "International Capital Flows and Portfolio Equilibrium," *Journal of Political Economy*, May/June 1974, 82, 443-67.
- Lucas, Robert E., Jr., "An Equilibrium Model of the Business Cycle," *Journal of Political Economy*, December 1975, 83, 1113-44.
- Mundell, Robert, "Capital Mobility and Stabilization Policy Under Fixed and Flexible Exchange Rates," *Canadian Journal of Economics and Political Science*, November 1963, 29, 475-85.
- Phelps, Edmund S. and Taylor, John B., "Stabilizing Powers of Monetary Policy under Rational Expectations," *Journal of Political Economy*, February 1977, 85, 163-90.
- Taylor, John B., "Conditions for Unique Solutions in Stochastic Macroeconomic Models with Rational Expectations," *Econometrica*, September 1977, 45, 1377-85.
- Turnovsky, Stephen J., "The Dynamics of Fiscal Policy in an Open Economy," *Journal of International Economics*, May 1976, 6, 115-42.

# The Kennedy Round: Evidence on the Regulation of International Trade in the United States

By HOWARD P. MARVEL AND EDWARD J. RAY\*

The purpose of this paper is to analyze and provide empirical evidence regarding the regulation of international trade flows. Our goal is to provide an explanation of the changes in the interindustry pattern of trade restrictions which emerged from the Kennedy Round of tariff negotiations. The principal conclusion of the analysis is that the pattern of protection which resulted from the Kennedy Round was shaped to minimize the domestic political disruption resulting from liberalization. Industries which faced significant import threats and whose characteristics enhanced their potential political influence were able to maintain tariff protection in the face of overall liberalization. The potential impact of the agreed-upon tariff reductions was partially undermined by the introduction of offsetting nontariff trade barriers (*NTBs*). In addition, *NTBs* were introduced systematically in industries that were becoming increasingly vulnerable to competition from imports. Indeed, our results suggest the possibility that the overall impact of the tariff reductions achieved and the concurrent erection of *NTBs* was to defeat in large measure the liberalizing impulse which motivated the Kennedy Round.

Section I outlines the institutional details of the Kennedy Round negotiations and presents an overview of the change in protection which resulted. This overview is followed by a brief summary of the economic theory of regulation and its implications for the regulation of international trade. Section II reports estimates of the determinants both of the interindustry pattern of Kennedy Round

tariff reductions and of the locus of *NTBs* that arose concurrently. The interindustry variation in tariff reductions appears to have depended on the ability of domestic industries to withstand enhanced import competition and on the characteristics of U.S. customers for the products in question. In particular, tariff reductions were relatively larger for rapidly growing domestic industries and smaller for consumer goods, relationships predicted by the economic theory of regulation. These politically induced modifications of the tariff reductions mandated by the Kennedy Round's linear rule were supplemented in industries particularly exposed to foreign competition by nontariff trade restrictions. Accordingly, the Kennedy Round tariff reductions cannot be interpreted independently of these *NTBs*. Section II contains estimates of their determinants that are used to test the extent to which *NTBs* were introduced to provide either supplementary protection to traditionally favored industries, compensatory protection to those industries that suffered substantial tariff reductions, or new protection to industries becoming increasingly vulnerable to import competition. Section III concludes with a brief assessment of our empirical findings.

## I. Kennedy Round Tariff Reductions and Industry Characteristics

A number of authors, including Robert Baldwin (1976) and John Cheh (1974), have investigated the implications of the Kennedy Round in terms of the cumulative benefits to consumers from access to cheaper imports and the short-run costs associated with industry and labor dislocation. Implicit in that literature is the assumption that the previously protected workers, industries, etc., lost their political power to obtain preferential trade restrictions and, as a consequence, con-

\*Ohio State University. We are grateful to William Cline, Charles Cox, Dale Larson, Robert Lipsey, Donald Parsons, Sam Peltzman, George Stigler, Thomas Wolf, participants in seminars at the Ohio State University and University of California-Davis, and the referees for helpful comments. Bruce Holloway and James Thomson provided excellent research assistance.

sumers would have access to cheaper goods from abroad. Using data for U.S. manufacturing in 1967, Marvel found that imports from abroad do restrain domestic industry profits (promote competitive pricing) and that imports are themselves quite sensitive to high domestic profits earned by U.S. manufacturing firms. Clearly, the removal of trade restrictions would be expected to enhance the ability of imports to promote competitive pricing in the U.S. market to the benefit of consumers in general.

The expressed purpose of the Kennedy Round was to reduce tariffs across the board by 50 percent—the so-called linear rule. The reductions were to be phased in beginning January 1, 1968, with the entire set of reductions to be in place by 1972. Measured against their overall objective, the negotiations were remarkably successful. The simple average tariff rate for U.S. manufacturing fell from 14.8 percent in 1965 to 11.4 percent in 1970, a year prior to full implementation of the negotiated reductions. Tariff rates computed by dividing duty collected by c.i.f. imports yield similar results: the simple average tariff reduction between 1967 and 1972 was 46.8 percent. Superficially, at least, imports seem to have responded to these reductions. The share of imports in domestic sales of manufactured products rose from an average value of 4.8 percent in 1967 to 7.3 percent in 1972, and the average annual growth rate in the value of domestic shipments fell substantially between 1967 and 1972 so that the average growth rate for the period 1958–72 of 5.7 percent was well below the 6.4 percent achieved for 1958–67. The impression one gets is that substantial tariff reductions permitted imports to rise rapidly in the United States and that imports displaced sales by less-competitive U.S. firms.

Upon closer investigation, the link between tariff reduction and increased imports becomes less apparent. Had the across-the-board liberalization envisioned in the Kennedy Round materialized, the largest cuts in absolute terms should have come in precisely those industries where high barriers had been erected to check potentially severe import competition. This pattern did not in fact obtain. The simple correlation between the

proportional change in duty-paid import prices due to Kennedy Round tariff cuts and the percentage change in imports across industries for the same period was 0.134, rather than negative and significant as a stimulative effect of cuts would predict.<sup>1</sup> Though the Kennedy Round reductions may well have increased access to the U.S. market, this lack of response suggests they were systematic rather than uniform across industries, and that substantial reductions in tariffs did not appear to have had any significant impact on the structure of imports into the United States.

We will argue that contrary to the presumption of earlier studies of the Kennedy Round, protected workers and industries did not lose their political power to obtain preferential treatment. The attempt to bypass domestic political and economic constraints through the imposition of negotiating rules proved unsuccessful. While trade in general may have been liberalized, the simple tariff reductions achieved overstate that liberalization. In our view, a primary function of the Kennedy Round in the United States was to redefine the structure of trade restrictions to maintain historic patterns of preferential protection and to promote the implementation of new forms of protection consistent with the requirements of changing market and political considerations.

The roots of our analysis can be found in the work of Sam Peltzman and Gary Becker. Peltzman argues that regulators do serve special interest groups, but they are constrained by the political risks associated with imposing costs on some of their constituents in order to help others. In the case at hand, trade restrictions provide rents to firms and workers in protected industries at the expense of domestic consumers in general, existing and/or potential foreign competitors,

<sup>1</sup>The simple correlation between the percentage change in tariff rates and the percentage change in imports was also insignificantly positive. The figure reported in the text is based on a referee's comment that large percentage cuts may have occurred in industries with low initial tariffs and low absolute tariff reductions. The text correlation employs as a tariff change measure  $(t_2 - t_1)/(1 + t_1)$ , where  $t_1$  and  $t_2$  are pre- and post-Kennedy Round tariff rates.

and indirectly, firms and workers in domestic industries with export potential. Changes in domestic and international demand and supply conditions, and in political activism by domestic consumers and foreign trading partners over time, are bound to lead policymakers to seek periodic changes in the structure of protection in order to reestablish a politically optimal menu of trade restraints. Becker makes the implicit argument for periodic reevaluation of regulation in his brief comment on Peltzman's paper. He argues that government intervention in a market can be usefully viewed as a politically optimal scheme to redistribute wealth from some constituents to others. If government regulation optimizes redistribution for given levels of political influence possessed by the parties to that redistribution and for given costs associated with the redistribution, changes in either the underlying political preference relations or the economic constraints on the redistributions will require a periodic restructuring of the regulatory scheme to reestablish a political equilibrium. In the present context, the liberalizing impulse reflected in the Kennedy Round's linear rule can be interpreted as disturbing a tariff structure in political equilibrium. The new tariffs which resulted are then a function both of the political power and economic interests of affected industries, and of the changes in industry characteristics occurring after the previous tariff regime had been established.

We will present evidence to suggest that the Kennedy Round tariff reductions, though perhaps increasing access to the U.S. market, were structured to minimize the political cost of import-induced domestic industry disruption. In addition, we will present evidence to suggest that nontariff trade restrictions were used systematically to offset losses in preferential treatment for domestic firms and workers that would otherwise have occurred as a result of tariff reduction.

## II. The Determinants of Tariff Reductions and Nontariff Barriers

Research by Baldwin (1971), Gary Hufbauer, and others on the determinants of U.S. imports and exports supports the con-

clusion that the United States possesses a comparative advantage in the production of research-intensive, skill-intensive, advanced technology products. Production of capital or low-skill-intensive products many of which are consumer durables, appears to take place at a comparative disadvantage. Richard Caves, Norman Fieleke, G. K. Helleiner, J. J. Pincus, Ray (1981a, b), Robert Stern, and Joe Stone have analyzed whether the pattern of tariff protection is responsive to comparative advantage, and each has produced evidence to support the view that tariff protection is most pronounced in industries susceptible to import incursions.

The concern of this paper is not with the pattern of protection at a point in time, but rather with the evolution of that pattern over time. Instead of focusing on considerations of dynamic comparative advantage, our concern is with the way in which political forces shape the inclusion of comparative advantage considerations into commercial policy. The studies cited above establish a link between comparative advantage and tariffs, but that link is not so close as to preclude political intermediation. Indeed, several of the variables customarily employed in such studies have obvious political interpretations. High levels of industrial concentration, for example, may facilitate the organization of cohesive lobbying efforts by reducing incentives to free ride.<sup>2</sup> Similarly, characteristics of a good's customers may condition their ability to exert pressure to forestall protection.

Our analysis focuses on four industry structure characteristics as proxies for the political influence (and political opposition) of industries anxious to affect their protection. The most salient of these is a measure of the domestic industry's recent strength—namely its growth rate. A central point of the Peltzman-Becker analysis of regulation is that windfalls or good fortune accruing to one party to a regulatory process will tend to be shared by offsetting regulations with the remaining parties. Similarly, poor performance by a regulated firm or industry will in some part be mitigated by

<sup>2</sup>See Mancur Olson for the classic discussion of free riding as a problem facing pressure groups.

transfers from that firm or industry's customers or rivals. Applied in the context of tariff policy, this result requires that rapidly expanding industries sacrifice some of their protection, hence economic rents, so that consumers may experience somewhat lower prices. "Sick," declining, or slow-growth industries will obtain increased protection as their consumers are forced by regulation to share in their ill fortune.<sup>3</sup> Hence the industry growth rate is expected to be negatively related to changes in tariff rates.

As noted above, market concentration and customer characteristics also are likely to influence an industry's ability to organize effective political pressure. Concentration plays a dual role during the Kennedy Round period. Not only did high levels of concentration facilitate organization to avoid tariff reductions, but also, the advent of non-tariff barriers provided an alternative to tariff protection that was more suited to the needs of low concentration industries. For both of these reasons, tariff cuts are expected to be most severe in low-domestic-concentration industries. Similarly, consumer goods producers with diffused opposition to protection are expected to have experienced relatively small tariff reductions.

One final variable is included in the analysis to take account of the requirement that the U.S. tariff reductions not only conform to U.S. political realities, but also aid in achieving offsetting tariff cuts and increased U.S. access to foreign markets. Since the United States appears to have been comparatively strong in high technology fields in the 1960's, and since such industries were likely candidates for "infant industry" protection overseas, it is expected that increased technological sophistication, as measured by the fraction of scientists and engineers in an

industry's workforce, will be associated with relatively large U.S. tariff reductions.

To test for these hypothesized relationships, regression estimates of the determinants of 1970 U.S. tariff rates were computed including the 1965 tariff rate among the explanatory variables. The data, with the exception of the consumer goods ratio, are from the U.S. International Trade Commission; the consumer goods ratio (personal consumption expenditure/total output) is from the 1972 Input/Output Study of the U.S. Economy. The data-mandated limitation to 1970 as the end of the period poses a slight problem in that the tariff rates employed do not reflect full implementation of the Kennedy Round cuts. Nevertheless, the phased tariff reductions, which had begun in 1968, were well reflected in the data by 1970.<sup>4</sup>

The regression results are reported in Table 1. The coefficient of the 1965 U.S. tariff rate is strongly significant, as expected. It is also significantly greater than 0.5, a result consistent with the state of the Kennedy Round's implementation in 1970. The point estimate of the coefficient, approximately 0.6, suggests that while the implementation was incomplete, it was nonetheless far enough along to permit testing of the political model of the negotiations. The results for the remaining four explanatory variables have the expected signs and differ significantly from zero.

The most striking result is the very strong relation between domestic industry growth and tariff reductions. The U.S. manufacturing industries experiencing rapid growth in employment were much more apt to have seen their tariff protection erode than were their less robust, and presumably more exposed, counterparts. The importance of the growth variable suggests strongly that the nominal tariff rate reductions achieved by the Kennedy Round overstate to a considerable degree the actual liberalization that re-

<sup>3</sup>This prediction is not so transparent as it might seem. If tariff regulations were simply of the "industry capture" variety, so that the tariff schedules were shaped by their influence on the profitability of directly affected firms, one would expect the highest tariffs to be obtained by rapidly growing firms. The benefits to such firms from reducing entry and protecting quasi rents would exceed those for declining firms, firms whose very unprofitability would both reduce outside competitive threats and discourage effective political cooperation.

<sup>4</sup>The 261 ITC industry observations in the sample were each weighted by the 1965 import share for that industry on the grounds that the negotiations would have focused most intensively on industries facing an important import threat while treating other industries as artifacts. The weighting did not markedly alter the estimates.

TABLE 1—DETERMINANTS OF 1970 U.S. MANUFACTURING TARIFF RATES

Variables	Coefficients <sup>a</sup> (Equations)			
	(1)	(2)	(3)	(4)
U.S. Tariff Rate, 1965	0.601 (21.97)	0.595 (22.22)	0.597 (22.12)	0.590 (22.46)
Four-Firm Concentration Ratio ( $C_4$ )	0.029 (2.19)	0.043 (3.15)	—	—
$\delta = \begin{cases} 0, & \text{if } C_4 < 0.5 \\ 1, & \text{if } C_4 \geq 0.5 \end{cases}$	—	—	1.99 (3.48)	2.59 (4.51)
Growth Rate (percentage change in total employment, 1958–67)	-5.19 (-4.46)	-4.78 (-4.18)	-5.50 (-4.80)	-5.04 (-4.50)
R&D Intensity (scientists and engineers as a percentage of total employment)	-17.06 (-4.24)	-15.60 (-3.95)	-18.00 (-4.54)	-16.38 (-4.23)
Consumer Goods Ratio	2.05 (2.08)	1.86 (1.93)	1.85 (1.90)	1.58 (1.66)
Percent of Production Workers in Unions	—	-0.040 (-3.65)	—	-0.042 (-4.03)
Intercept	6.34 (4.34)	7.74 (5.23)	7.33 (5.21)	9.28 (6.40)
$R^2$	.78	.78	.78	.80

Source: Data from U.S. International Trade Commission except as noted in text.

Note: 261 observations, weighted by 1965 import share.

<sup>a</sup>t-values are shown in parentheses.

sulted from their implementation. Each of the remaining coefficients also contributes support to the view that structural characteristics of U.S. manufacturing shaped the tariff reductions that obtained. High-concentration industries were relatively successful in organizing effective opposition to tariff reductions, though as we shall see below, some of this result may be due to the comparative attractiveness of nontariff barriers as a tool for protection in low-concentration industries. Organized opposition seems also to have affected the outcome of the negotiating process. Industries selling primarily to consumers were better able to avoid tariff reductions than were those who marketed intermediate products. The only segment of manufacturing that appears to go against this trend of mitigating the Kennedy Round tariff cuts consisted of high-technology products. As we have argued above, those cuts were likely to have been accepted primarily to achieve parallel reductions in foreign tariff rates.

The estimates of the determinants of 1970 tariff rates have concentrated on industry structure characteristics as explanatory vari-

ables and have thereby focused on business as the source of political influence on the negotiating process. This may seem too narrow in light of the substantial publicity received by various union efforts to stem foreign competition and convince U.S. consumers to "buy American." There is, in addition, substantial reason to suggest that union attitudes toward protection had shifted markedly in the period prior to the Kennedy Round negotiations. Alan Deardorff and Robert Stern catalog the attitude of American labor toward free trade by reference to congressional voting patterns and conclude that while protectionist sentiment may have abated slightly during the 1960's, its trend was to increase during the entire postwar period. It therefore seemed reasonable to include union influence on the Kennedy Round negotiations in the regression context. Union strength is measured by the percentage of an industry's production workers who are union members.<sup>5</sup> The results (Table

<sup>5</sup>The data are from Richard Freeman and James Medoff.

TABLE 2—NONTARIFF BARRIERS IN U.S. MANUFACTURING, 1970

Variables	Coefficients <sup>a</sup> (Equations)			
	(1)	(2)	(3)	(4)
U.S. Tariff Rate, 1965	0.025 (2.33)	0.023 (2.12)	0.025 (2.35)	0.023 (2.11)
Change in Tariffs, 1965–70 (percentage points)	0.025 (1.94)	0.024 (1.85)	0.026 (2.00)	0.025 (2.11)
Four-Firm Concentration Ratio ( $C_4$ )	–0.011 (–2.47)	–0.010 (–2.35)	–	–
$\delta = \begin{cases} 0, & \text{if } C_4 < 0.5 \\ 1, & \text{if } C_4 \geq 0.5 \end{cases}$	–	–	–0.53 (–2.62)	–0.051 (–2.54)
Consumer Goods Ratio	0.779 (2.76)	0.776 (2.74)	0.825 (2.91)	0.820 (2.88)
Percent of Production Workers Unionized	–	–0.0026 (–0.70)	–	–0.0030 (–0.80)

Source: Data from U.S. International Trade Commission except as noted in text.

Note: Dependent Variable = 1, if any nontariff barrier present (69 industries) and 0, otherwise (192 industries). Probit estimates, 261 TCSIC industries.

<sup>a</sup>Asymptotic *t*-values shown in parentheses.

1, equations (2) and (4)) are surprising. Heavily unionized industries experienced significantly greater tariff reductions than did their nonunion counterparts. Though one certainly cannot attribute causation to this union effect, it does suggest that organized labor simultaneously recognized the threat to its membership by the Kennedy Round while it either did not exert or failed to possess the political pressure required to derail or at least to reshape the cuts.

These tariff regressions present an informative but incomplete picture of the way in which the Kennedy Round ground rules were circumvented. Nontariff barriers were available as a device either to offset the impact of potentially painful tariff reductions, or to augment protection in industries where foreign competition was becoming troublesome. The remainder of this section deals with the interindustry incidence of nontariff protection as of 1970. Though 1964 *NTB* data are unavailable, we treat the *NTB* estimates analogously to those for 1970 tariffs on the grounds that growth of *NTBs* in the late 1960's was so rapid that it seems appropriate to treat the *NTBs* as first available for widespread application in manufacturing at the time of the Kennedy Round.

In general the interindustry locus of *NTBs* seems to be determined by forces similar to

those shaping the pattern of tariff protection (Ray, 1981b), so that tariffs and *NTBs* are largely complementary. This complementarity does not preclude the possibility that *NTBs* were imposed as substitutes for tariff protection sacrificed in the Kennedy Round. If the Kennedy Round's linear rule had served effectively to constrain the ability of U.S. regulators to shape the tariff structure in response to political pressure, *NTBs*—not subject to the negotiations—could have been used to offset undesirably large tariff reductions. On the other hand, if the Kennedy Round rules did not prove particularly burdensome in practice, the *NTBs* could have been imposed in industries which required not only avoidance of tariff reductions, but actual increases in protection. Were this the case—that is, if *NTBs* complemented rather than substituted for tariff modifications—it would suggest that the attempt to constrain the negotiations through the linear rule was even less successful than indicated by our tariff regressions.

To test whether *NTBs* and the Kennedy Round tariff modifications were complements or substitutes, a dummy variable taking the value of one in the presence of at least one *NTB* applying to an industry was regressed on the 1965 tariff rate, the change in rates occurring between 1965 and 1970,

and a set of other explanatory variables described below. A positive coefficient on the tariff-change variable would support the complementary view of *NTBs* while a negative value would suggest they were used to offset Kennedy Round tariff reductions. The results in Table 2 support the view that the *NTBs* actually augmented protection in favored industries.

Though tariffs and *NTBs* may, on balance, be complementary, it is clear that the two forms of protection may differ markedly in their attractiveness for specific industries. In particular, *NTBs* are apt to be relatively more accessible than tariffs for low-concentration industries. These industries can be expected to experience greater difficulties in generating political influence than their high-concentration counterparts because of free rider and other policing problems involved with this sort of cooperative effort. Moreover, any rents generated by tariff protection obtained with such political influence are more susceptible to being bid away through rapid entry of domestic firms into a low-concentration industry. Many nontariff barriers generate measurable benefits in excess of those which tariffs provide, rents which can be used to induce membership in a lobbying arrangement by either existing recalcitrant firms or potential opponents. These rents can often be distributed selectively to punish free riders, whether they be exiting firms or new entrants. For example, quotas based on historic sales can be used to distribute the benefits of trade restraint to all domestic holders of industry-specific factors of production. Voluntary export restraints distribute the benefits to foreign producers, and in that way reduce international political pressure for freer trade. These advantages of nontariff barriers should be compared to their costs to the industry involved. By broadening the base of support for a particular governmental action, one also draws into the regulatory process other interests with differing and occasionally conflicting goals. Therefore, were the difficulties of obtaining tariff and nontariff barriers identical, tariff protection would be preferred by the protected industry. But those difficulties are not the same, and therefore the advantages of nontariff

barriers in generating political support make it possible for low-concentration industries to obtain nontariff barriers to trade even though they lack the political power to prevent substantial tariff cuts.

These considerations suggest that to the extent that substitution between *NTBs* and tariffs occurs, low-concentration industries should be more apt to seek the former. The regression results confirm this. While high-concentration industries were relatively effective in maintaining existing tariff protection (Table 1), *NTBs* are significantly more likely in low-concentration industries. To the extent that such industries were previously unable to exert pressure to obtain tariffs, this finding may suggest protection is becoming more pervasive.

The remaining variable, the consumer goods ratio, suggests that *NTBs* are relatively likely to arise in consumer goods industries. This may reflect the lack of political power on the part of consumers and their representatives. Alternatively, quota rights or exemptions from regulation may be assigned to established distributors to forestall their opposition to *NTBs*. In either case, the result is consistent with the tariff equation finding of increasing effective protection of downstream producers.<sup>6</sup>

### III. Summary and Conclusions

Our findings suggest that despite the attempt to limit negotiation discretion by the imposition of prior rules, political pressure appears to have shaped the pattern of protection which emerged from the Kennedy Round. It is inappropriate to interpret our argument as suggesting that the free trade movement cannot prevail generally, or that it did not prevail in the Kennedy Round. We simply suggest that special interests work systematically to undermine the trade liberalization where its effects would be most

<sup>6</sup>The specifications of the *NTB* estimates also include a unionization variable for consistency with the tariff estimates. Surprisingly, unionization does not appear to affect the locus of *NTBs* at all. It is clear that any defeats suffered by unions on tariff issues were not offset by *NTB* imposition.

painful politically. Our empirical results provide preliminary evidence regarding the industrial characteristics of industries that appear to be relatively successful in withstanding external pressure to permit potential foreign competitors to enter their domestic markets.

## REFERENCES

- Baldwin, Robert E., "Determinants of the Commodity Structure of U.S. Trade," *American Economic Review*, March 1971, 61, 126-46.
- , "Determinants of the Commodity Structure of U.S. Trade: Reply," *American Economic Review*, June 1972, 62, 465.
- , "Trade and Employment Effects in the United States of Multilateral Tariff Reductions," *American Economic Review Proceedings*, May 1976, 66, 142-48.
- Becker, Gary, "Comment," *Journal of Law and Economics*, August 1976, 19, 245-48.
- Caves, Richard E., "Economic Models of Political Choice: Canada's Tariff Structure," *Canadian Journal of Economics*, May 1976, 9, 278-300.
- Cheh, John H., "United States Concessions in the Kennedy Round and Short-Run Labor Adjustment Costs," *Journal of International Economics*, November 1974, 4, 323-40.
- , "A Note on Tariffs, Nontariff Barriers and Labor Protection in U.S. Manufacturing Industries," *Journal of Political Economy*, April 1976, 84, 389-94.
- Deardorff, Alan V. and Stern, Robert M., "American Labor's Stake in International Trade," in *Tariffs, Quotas and Trade: The Politics of Protectionism*, San Francisco: Institute for Contemporary Studies, 1979, 125-48.
- Fieleke, Norman S., "The Incidence of the U.S. Tariff Structure on Consumption," *Public Policy*, Fall 1971, 19, 639-52.
- Freeman, Richard B. and Medoff, James L., "New Estimates of Private Sector Unionism in the United States," *Industrial and Labor Relations Review*, January 1979, 32, 143-74.
- Helleiner, G. K., "The Political Economy of Canada's Tariff Structure: An Alternative Model," *Canadian Journal of Economics*, May 1977, 10, 318-26.
- Hufbauer, Gary C., "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods," in Raymond Vernon, ed., *The Technology Factor in International Trade*, New York: National Bureau of Economic Research, 1970, 145-231.
- Marvel, Howard, P., "Foreign Trade and Domestic Competition," *Economic Inquiry*, January 1980, 18, 103-22.
- Olson, Mancur, Jr., *The Logic of Collective Actions: Public Goods and the Theory of Groups*, New York: Schocken, 1968.
- Peltzman, Sam, "Toward a More General Theory of Regulation," *Journal of Law and Economics*, August 1976, 19, 211-40.
- Pincus, J. J., "Pressure Groups and the Pattern of Tariffs," *Journal of Political Economy*, August 1975, 83, 757-78.
- Ray, Edward John, (1981a) "The Determinants of Tariff and Nontariff Trade Restrictions in the United States," *Journal of Political Economy*, February 1981, 89, 105-21.
- , (1981b) "Tariff and Nontariff Barriers to Trade in the United States and Abroad," *Review of Economics and Statistics*, May 1981, 63, 161-68.
- Stern, Robert M., "The U.S. Tariff and the Efficiency of the U.S. Economy," *American Economic Review Proceedings*, May 1964, 54, 459-70.
- Stone, Joe A., "A Comment on Tariffs, Nontariff Barriers, and Labor Protection in U.S. Manufacturing Industries," *Journal of Political Economy*, October 1978, 86, 959-62.

# The Rational Expectations Hypothesis in Retrospect

By FILIPPO CESARANO\*

The rational expectations hypothesis is at the center of current debates in economic theory. The importance of this theoretical innovation has stimulated a number of studies, including analyses in the unfashionable field of the history of thought, the 1979 article by Brian Kantor being a first example. Historical investigations on the subject of rational expectations are of interest since it is held that this theory has revived the teachings of the classics, inasmuch as it provides the cornerstone of the recently developed approach known as "new classical macroeconomics." The present note examines some specific aspects, as yet not emphasized in the literature, concerning the relationship between rational expectations and classical monetary theory. In particular, the following results will be shown: the fundamental implication of the rational expectations hypothesis, that is, that only unexpected changes in the money stock influence the real sector of the economy, is a basic tenet of classical monetary theory as well and is founded upon the very argument according to which agents efficiently use the available information in order to take account of the consequences of policy measures (Section I). This common analytical framework notwithstanding, classical analysis of the role to be assigned to the monetary authority reflects a different appraisal of the dimension of the information set (Section II).

## I

Classical economists have discussed at length the effects of both unexpected and expected changes in the money stock, but have paid far greater attention to the former. In the monetary system of the time, that is, a

metallic standard, variations in money supply essentially originated from the balance of trade and from the discovery of new mines. Thus, given the lack of regularly collected statistics, the classics normally assumed such variations to be unexpected.<sup>1</sup> Debasement of the currency by the prince, the other possible source of variation in the nominal money stock, represented, under certain conditions, the case of expected change in money, but was regarded more as an exceptional measure than an ordinary course of action.

Classical economists viewed the transmission mechanism of unexpected monetary impulses as a sequential process determined by the gradual propagation of the increase in money in the economy. It takes time for this sequential process to be completed and, in the transition, real income and employment are affected before the price level adjusts. The *locus classicus* of this mechanism is a

<sup>1</sup>In his thorough examination of the subject, Richard Cantillon notes: "It is also usually the case that the increase or decrease of actual money in a State is not perceived because it flows abroad, or is brought into the State, by such imperceptible means and proportions that it is impossible to know exactly the quantity which enters or leaves the State" (1755, p. 163). The assumption was also discussed in hypothetical terms. This clearly appears in the following passage in which John Locke describes the effects of a reduction in the quantity of money, focusing on the notion of an equilibrium amount of cash balances:

"If one-third of the money employed in trade were locked up, or gone out of England, must not the landholders necessarily receive one-third less for their goods, and consequently rents fall; a less quantity of money by one-third being to be distributed amongst an equal number of receivers? Indeed, *people not perceiving the money to be gone*, are apt to be jealous one of another; and each suspecting another's inequality of gain to rob him of his share, every one will be employing his skill and power the best he can to retrieve it again, and to bring money into his pocket in the same plenty as formerly." [1691, pp. 70-71, emphasis added]

The distinction should be made between *unexpected* and *unperceived* changes in the money stock. Such a distinction, however, does not significantly bear upon the main arguments of the present paper.

\*Banca d'Italia. A first draft of this paper was written while I was a visiting scholar at Harvard University. Helpful comments from Benjamin Friedman, David Laidler, and the referee are gratefully acknowledged. The usual disclaimer applies.

celebrated passage by David Hume,<sup>2</sup> who clearly identifies the basic argument accounting for the short-run effectiveness of monetary impulses. Through the several stages of the transmission mechanism, agents have no information about the change in the money stock and, thus, do not modify their behavior accordingly. Hume states:

When any quantity of money is imported into a nation, it is not at first dispersed into many hands, but is confined to the coffers of a few persons, who immediately seek to employ it to advantage.... They are thereby enabled to employ more workmen than formerly, *who never dream of demanding higher wages*, but are glad of employment from such good paymasters.... It is easy to trace the money in its progress through the whole commonwealth; where we shall find, that it must first quicken the diligence of every individual, before it encrease the price of labour.

[1752, p. 38, emphasis added]

It should be stressed that, in analyzing the efficacy of changes in money, classic authors did not refer to peculiar assumptions relating to institutional features, such as rigid prices or long-term contracts. They focused, instead, upon the lack of information during the time necessary for the mechanics of the increase in money to work its effects through the economy.<sup>3</sup> Indeed, the world of the

classics is one of flexible wages and prices where long-term contracts do play a role but account only for the redistributive effects of inflation (for example, see Cantillon, 1755, pp. 163–65). Thus, it is not price or wage rigidity that distinguishes classic writers from rational expectations theorists on this specific point, but the particular element which is missing in the information set: the former consider the unavailability of information on the change in the money stock, which does not allow individuals to shift their behavior functions; the latter look at the lack of data on the current price level, which does not enable private agents to assess the relative prices of their outputs.<sup>4</sup>

A rudimentary version of the second hypothesis can nevertheless be found in Ferdinando Galiani's analysis of the effects of an "augmentation," that is, a once-and-for-all increase in the nominal quantity of money implemented through currency debasement. He puts forward a theory of the agents' "connection of ideas about the prices of commodities and money," according to which individuals conceive their "ideas" about prices from "truth," that is, from the

---

ble not to lose sight of it seeing that having been amassed to make large sums it is distributed in the little rills of exchange, and then gradually accumulated again to make large payments." [1755, pp. 161–63]

The discussion of the transmission mechanism by Cantillon (pp. 159–99) is far more detailed and, it should be stressed, predates Hume's *Discourses* since the *Essai* was actually written sometime in the 1720's or early 1730's. In particular, Cantillon attempts to show that: (i) the effectiveness of variations in money supply depends on the channel through which money is injected (this has been called the "Cantillon Effect" by Mark Blaug, 1978, p. 22); (ii) the price level does not necessarily change in proportion to the variation in the money stock; (iii) the structure of relative prices is also affected.

<sup>4</sup>Therefore, it does not seem entirely correct to contrast the rational expectations theorists' "information lag hypothesis," which assumes flexible wages and prices, with the classics' "adjustment lag hypothesis" related "at least partially to the existence of long-term contracts" (Charles Nelson, 1981, pp. 1–2). A not-inconspicuous consequence of these interpretation failures is to attribute a Keynesian character to the work of Milton Friedman and of the St. Louis economists on the grounds that both postulate a slowly adjusting price level in the transmission mechanism (see Nelson, pp. 3–4).

<sup>2</sup>"...[W]e must consider, that though the high price of commodities be a necessary consequence of the encrease of gold and silver, yet it follows not immediately upon that encrease; but some time is required before the money circulates through the whole state, and makes its effect be felt on all ranks of people. At first, no alteration is perceived; by degrees the price rises, first of one commodity, then of another; till the whole at last reaches a just proportion with the new quantity of specie which is in the kingdom. In my opinion, it is only in this interval or intermediate situation, between the acquisition of money and rise of prices, that the encreasing quantity of gold and silver is favourable to industry."

[1752, pp. 37–38]

<sup>3</sup>Cantillon notes: "I have already remarked that an acceleration or greater rapidity in circulation of money in exchange, is equivalent to an increase of actual money up to a point.... On the other hand money flows in detail through so many channels that it seems impossi-

true state of things; and, to those "ideas," they pair the "sound of names," that is, actual or called prices. (See Galiani, 1751, pp. 68–72.) Augmentation does not affect the true state of things but only actual prices. The latter, however, do not change immediately since agents take some time to get accustomed to higher prices. It is this delayed response which allows the monetary authority to gain from an inflationary measure. Galiani states:

Augmentation of money is a profit that the prince and the state make from the slowness with which the multitude changes the connection of ideas about the prices of goods and money....It [augmentation] does not produce any change in things, but in names; thus in order for the prices of things to remain the same, they have also to change with regard to their names. If this happened on the same day in which the augmentation was carried out and happened entirely, and entirely in proportion, then the augmentation would have no consequences; just as a law which established that money, instead of being called by Italian names, should be called by Latin or Greek or Hebrew names would have no consequences.

[pp. 68–70]

It could be pointed out that, essential to the theory of rational expectations, is the inclusion in the agents' information set of the "true" model of the economy. This is, without doubt, a novel assumption in the modelling of individual behavior that has not been set forth in earlier writings. In the historical reconstruction of economics, however, what is to be investigated are not precise and formal descriptions of a theory—since this would be a preposterous task—but rather the essential concepts underlying them. It can be shown that the classics commonly adopted the assumption, albeit failing to state it explicitly, according to which agents correctly anticipate the implications of announced decisions by the monetary authority with the help of the relevant theoretical framework. In fact, this is a necessary condition for those correct im-

plications to be derived: this obtains only if agents do make use of a theory which links together the available data. The remainder of this section provides evidence for the latter proposition.

As explained earlier, the classics gave less weight to discussions of expected variations in the quantity of money. Nevertheless, in those cases in which a change in money supply was deliberately engineered by the prince,<sup>5</sup> and publicly announced, classic authors emphasize the consequent shift in the agents' behavior, thereby accounting for the inefficacy of such policies. Cantillon (1755, p. 289 ff.) refers to the deflationary measure implemented by the king of France in 1714 when, by decree, the nominal value of the *écu* was decreased from 5 to 4 *livres*, 1 percent per month for a twenty-month period. He describes in detail the implications of this measure and, notwithstanding the inconsistency of some of the arguments, shows that the agents responded to an announced policy as if they understood the correct model. Hence, their behavior is not policy invariant.<sup>6</sup>

Hume was also aware of the inefficacy of expected monetary impulses. Therefore, while proposing an inflationary measure, he pur-

<sup>5</sup>An increase in the money stock could be implemented in three ways: by decree, giving a new face value to the coins without altering their metallic content; by melting and coining again the money stock, but lowering the titer of each coin; by cutting a hole in the middle of the coins.

<sup>6</sup>Cantillon states:

"Enlightened people in France hoard their money in these times. The King finds means to borrow much money on which he willingly loses the diminution, proposing to compensate himself by an augmentation at the end of the diminution. With this object after several diminutions they begin to hoard money in the King's Treasury, to postpone the payments, pensions, and army pay. In these circumstances money becomes extremely rare at the end of the diminutions both by reason of the sums hoarded by the King and various individuals and by reason of the nominal value of the coin, which value is diminished." [1755, pp. 289–91].

While explaining the failure of the king of France to gain from the augmentation of 1726, Cantillon again establishes that individuals react differently to unexpected policies: "The diminutions which had preceded this augmentation were made suddenly without warning, which prevented the ordinary operations of diminutions" (p. 295).

posely wished it to be a "surprise."<sup>7</sup> Finally, Galiani's analysis stems from his theory of the "connection of ideas about the prices of commodities and money." Once the information on the monetary shock is available, individuals promptly revise their "ideas" on prices and the latter increase without lags: the inflationary policy becomes ineffective.<sup>8</sup>

## II

In Section I, it was shown that the proposition concerning the efficacy of unexpected monetary impulses, advanced by the classics, is founded upon the very principle of rational behavior on the part of agents, characteristic of modern monetary theory. It may be natural, then, to ask which implications for the conduct of the monetary authority were derived from this common analytical groundwork.

In a commodity standard, the scope for the authority's action is constrained by the very rules of the monetary system. Hence, with regard to the works of the classics, it is not appropriate to refer to "monetary policy" as in relation to modern writings, since the role of the authority in the respective institutional frameworks, a commodity standard on the one side and a fiat-money system on the other, can hardly be compared. Indeed, the classics could not have arrived at a full-fledged conception of policy design, linking instruments to targets, since the prerequisites for such a conception were missing. This impossibility of comparison notwithstanding, it is of interest to inquire about the classics' analyses of the authority's role inasmuch as this can be suggestive of a better understand-

ing of the relationship between classical and contemporary monetary theory.

Currency debasement was within the prince's range of power, but was commonly regarded as an interference with the smooth working of a commodity standard. Even those few authors who spoke favorably of debasement were careful to qualify their position. Galiani, for example, does not regard "augmentation" of money as an ordinary measure to be resorted to in normal times, but as an exceptional one to be taken only in emergencies. The monetary authority should evaluate the benefits and costs of such a measure and implement it whenever the net effect is positive (1751, p. 68). This is likely to be the case when the prince faces a critical situation and cannot raise taxes or issue securities or sell his own property. In these circumstances, augmentation is the right course of action since its effects will be slow and will be spread over the entire population (see Galiani, pp. 87-117).

The negative attitude of the classics towards debasement was related not only to the view of debasement as a disruption of the monetary arrangements, but also to its characteristic of being often a once-and-for-all change as well as an expected change in the money stock. On the contrary, variations in the quantity of money inherent in the functioning of a metallic standard (for example, those originating from the balance of trade and from the discovery of new mines) were both continuous and unexpected and, thus, were thought to influence real variables for a not ephemeral period. Cantillon, aware of the effects on output stemming from balance of trade adjustment, does advocate a norm of conduct designed to control the quantity of money in order to prevent economic fluctuations:

When a State has arrived at the highest point of wealth (I assume always that the comparative wealth of States consists principally in the respective quantities of money which they possess) it will inevitably fall into poverty by the ordinary course of things. The too great abundance of money, which so long as it lasts forms the power of States, throws them back imperceptibly but

<sup>7</sup>"In executing such a project [taking from every shilling a penny's worth of silver], it would be better to make the new shilling pass for 24 halfpence, in order to preserve the illusion, and make it be taken for the same" (1752, p. 39, emphasis added).

<sup>8</sup>Galiani states:

"If he [the prince] abused it [augmentation], the connection collapses, names change their significance, things remain the same and the unsurpassable force of nature wins.... Finally, a prince who, abusing augmentation, practised it every month, destroying every connection of ideas between prices and commodities, would render it quite useless and inefficacious; and only with other policies would obtain what today is obtained with augmentation." [1751, pp. 70-71]

naturally into poverty. Thus it would seem that *when a State expands by trade and the abundance of money raises the price of Land and Labour, the Prince or the Legislator ought to withdraw money from circulation, keep it for emergencies, and try to retard its circulation by every means except compulsion and bad faith, so as to forestall the too great dearness of its articles and prevent the drawbacks of luxury.*

[1755, p. 185, emphasis added]

In the subsequent statement, however, Cantillon recognizes the difficulties of implementing such a stabilizing action: "...it is not easy to discover the time opportune for this, nor to know when money has become more abundant than it ought to be for the good and preservation of the advantages of the State" (p. 185).

In an earlier passage, he also points out the scanty knowledge about the properties of the transmission mechanism:

The proportion of the dearness which the increased quantity of money brings about in the State will depend on the turn which this money will impart to consumption and circulation. Through whatever hands the money which is introduced may pass it will naturally increase the consumption; but this consumption will be more or less great according to circumstances. It will be directed more or less to certain kinds of products or merchandise according to the idea of those who acquire the money....I conceive that when a large surplus of money is brought into a State the new money gives a new turn to consumption and even a new speed to circulation. But it is not possible to say exactly to what extent.[pp. 179-81]

Cantillon's recommendation for controlling the time path of the money stock parallels, keeping in mind the caveat stated at the beginning of this section, the policy rule called for by both Friedman and rational expectations theorists. This very rule of conduct, however, is founded upon quite distinct arguments. According to Milton Friedman, a

discretionary monetary policy is likely to increase rather than decrease instability, because of the insufficient knowledge of the transmission mechanism, together with the presence of long and variable lags. According to rational expectations theorists, instead, private agents, and a fortiori the monetary authority, have complete knowledge of the "true" model of the economy; a discretionary policy will increase the variance of output because, by conveying wrong signals to agents, it blurs the assessment of relative prices. Hence, there seems to be a clear-cut divergence between the rational expectations theorists' and Friedman's standpoint concerning the dimension of the information set. All in all, it is Friedman's work which appears to be closer to the classical tradition.<sup>9</sup>

Following the message of this short note, much care should be taken in representing the rational expectations literature as a punctilious continuation of the teachings of classical monetary theory. There are, in fact, some conspicuous asymmetries and their study may well provide stimulating ideas for the further advancement of monetary economics.<sup>10</sup>

<sup>9</sup>See especially where Friedman states: "As I see it, we have advanced beyond Hume in two respects only: first, we now have a more secure grasp on the quantitative magnitudes involved; second, we have gone one derivative beyond Hume" (p. 177).

<sup>10</sup>The role the history of economic analysis can play in furthering current research is thoroughly investigated in my 1983 article.

## REFERENCES

- Blaug, Mark, *Economic Theory in Retrospect*, New York: Cambridge University Press, 1978.
- Cantillon, Richard, *Essai sur la Nature du Commerce en Général*, London, 1755 (Henry Higgs, ed., London: Macmillan, 1931).
- Cesarano, Filippo, "On the Role of the History of Economic Analysis," *History of Political Economy*, Spring 1983, 15.
- Friedman, Milton, "25 Years After the Rediscovery of Money: What Have We Learned? Discussion," *American Economic Review*

- Proceedings*, May 1975, 65, 176-79.
- Galiani, Ferdinando, *Della Moneta*, Naples, 1751; reprinted in Pietro Custodi, ed., *Scrittori Classici Italiani di Economia Politica*, Vol. IV, Milan: Destefanis, 1803.
- Hume, David, "Of Money," in *Political Discourses*, Edinburgh, 1752; reprinted in Eugene Rotwein, ed., *Writings on Economics*, Toronto: Nelson, 1955.
- Kantor, Brian, "Rational Expectations and Economic Thought," *Journal of Economic Literature*, December 1979, 17, 1422-41.
- Locke, John, *Some Considerations of the Consequences of Lowering the Interest and Raising the Value of Money*, 1691; reprinted in *Works*, Vol. V, London, 1823.
- Nelson, Charles R., "Adjustment Lags Versus Information Lags," *Journal of Money, Credit, and Banking*, February 1981, 13, 1-11.

# The Distributional and Efficiency Effects of Increasing the Minimum Wage: A Simulation

By WILLIAM R. JOHNSON AND EDGAR K. BROWNING\*

The generally accepted goal of minimum wage laws is to alter the distribution of income in favor of low-income households. However, since a minimum wage will also disrupt low-wage labor markets, causing inefficiencies and imposing costs on some of the same workers the law is intended to help, an evaluation of the policy requires a weighing of the distributional benefits against the efficiency costs. Quantification of these benefits and costs is clearly important to this evaluation. While much effort has been devoted to estimating the employment effects of minimum wage laws, relatively little work has been done to estimate the distributional benefits. Our simulations focus on this efficiency-equity tradeoff by developing estimates of the impact of an increase in the minimum wage on the level and distribution of real income across households.

Previous work has concluded that minimum wages are not as beneficial in their distributional impact as generally supposed.<sup>1</sup> Basically, this results from two important facts about the structure of household income. First, low-wage workers are frequently members of high-income households, so a large part of any increased wages produced by the policy will accrue to the upper part of the income distribution. Second, low-income households receive a relatively small share of their income from low-wage earnings so that even large increases in the earnings of low-wage workers do not cause large increases in the income of low-income households. The simulations reported herein build on this

previous research and extend it in several ways: 1) the micro data base employed in this study contains estimates of most government transfers and taxes, allowing us to estimate the impact of an increase in the minimum wage on the distribution of net (disposable) income; (2) estimates of the distributional impact of the costs an increase in the minimum wage imposes on the general public are included; and (3) the present study takes into account the fact that higher wage income produced by minimum wage policy will result in lower transfers received and higher taxes paid by low-income households, thereby diminishing the gains in net income in proportion to the beneficiaries' effective marginal tax rates.

Section I describes the simple theoretical model which underlies our calculations and considers some of the characteristics of the data that are relevant for this study. Section II contains the major results of the study; it reports estimates of the distributional and efficiency impact of the minimum wage on the 1976 distribution of income among households, under a variety of alternative assumptions. Section III contains some concluding observations.

## I. Theoretical Framework and Data

The basic model used for our calculations is a competitive model of the low-wage labor market in which all other factors of production (high-wage labor, capital, etc.) are treated as a composite input which is in perfectly inelastic supply. The computations to be presented embody three assumptions, the relaxation of which, we believe, would not appreciably alter the tenor of our results. The first assumption, that all other factors of production can be considered as a composite, essentially precludes any "spillover" effects of the minimum wage on labor not

\*Department of economics, University of Virginia. This research has been supported by the Minimum Wage Study Commission. The views expressed are solely our own, and do not necessarily represent the views of the Commission. We thank Robert Goldfarb, Daniel Saks, and Jacqueline Browning for helpful comments.

<sup>1</sup>See, for example, the studies by Edward Gramlich and Thomas Kniesner.

directly affected by the minimum and implies that the cost of a higher minimum will be borne proportionately by all other factors.<sup>2</sup>

The second assumption is that the minimum wage covers all market work so that the relevant labor supply curve is the supply of low-wage labor to the labor market rather than the supply to a smaller covered sector. This assumption is justified by the small size of the uncovered sector and the theoretical ambiguity of two-sector models.<sup>3</sup> Third, a one-product economy is postulated, which allows us to ignore the general equilibrium effects of the changes in relative product prices on the distribution of income.

Two complications of the simple model described above, however, must be addressed by the simulations. First, low-wage workers are not homogeneous, but will have wages ranging up to the new minimum wage. We assume that the elasticity of substitution between all grades of low-wage labor is infinite, so that the same elasticity of demand applies to all low-wage workers. Hence the reduction in employment of workers at each wage level is proportional to the increase in their wages caused by the increase in the minimum wage.

The second complication accounts for the considerable marginal tax rates of both the tax and transfer systems which temper both the income gains of low-wage workers and the income losses of other resource owners. In the zero demand elasticity case, this means that the sum of gains and losses to resource owners equals the change in the government budget deficit.

We developed the data base used in this study for the purpose of investigating the distribution of the tax burden by income classes.<sup>4</sup> Let us consider the distribution

across household income classes of those low-wage workers directly affected by an increase in the 1976 minimum wage from \$2.30 to \$2.80 (the simulation we perform). As Gramlich and others have emphasized, the relation between low-wage rates for individual workers and low household income is much weaker than commonly believed. Many low-income households have no workers at all (for example, the retired and many female-headed households), while high-income households frequently have several workers, among whom are likely to be some secondary workers earning relatively low wages. Table 1 shows the significance of these factors for an increase in the minimum wage from \$2.30 to \$2.80. While high-wage workers are disproportionately found in high-income households, Table 1 indicates that low-wage workers are about evenly distributed across the income distribution. In fact, slightly more than half of all low-wage workers are members of households in the upper half of the income distribution.

Not only are low-wage workers not concentrated in low-income households, 40 percent of workers in the bottom income decile have wage rates in excess of \$2.80. Furthermore, column (4) in Table 1 indicates the unimportance of low-wage labor earnings relative to total after-tax, after-transfer (*ATAT*) income for each decile. Since more than 86 percent of the income of the lowest decile derives from sources other than low-wage work, it is clear that increasing the minimum wage will have a modest effect on total income.

One of the features of this study is an emphasis on the fact that any gains in wages

<sup>2</sup>In our 1981 paper we relaxed this assumption to allow spillover effects without appreciable consequences.

<sup>3</sup>Finis Welch states that "...the 'full coverage' model is probably the appropriate one for describing today's effects" (p. 28).

<sup>4</sup>See our 1979 and 1981 studies for a complete description of the data. Wage rates and the number of hours worked for all workers in the population are not readily available from the March *Current Population Survey*. We calculated the wage rate (or average hourly earnings, more precisely) for the past week by dividing

last week's earnings by last week's hours of work and assumed that labor earnings for the past year were earned at this wage rate to compute last year's hours of work. Because of missing or obviously incorrect data, there are a sizeable number of workers for which wage rates cannot be computed. As we argue in greater detail in our 1981 paper, the bias introduced by neglecting these workers may be somewhat larger for the lower-income classes. Although the extent of this bias cannot be determined without knowledge of the distribution of true wage rates for the group with missing data, our estimates suggest it is not likely to be large enough to substantially affect the results.

TABLE 1—DISTRIBUTION OF WORKERS BY WAGE RATES ACROSS HOUSEHOLD INCOME AND IMPORTANCE OF LOW-WAGE WORK (Shown in Percent)

Income Decile	Workers with Wage $\leq$ \$2.30 (1)	Workers with Wage $>$ \$2.30 and $\leq$ \$2.80 (2)	Workers with Wage $>$ \$2.80 (3)	Earnings of Workers with Wage $\leq$ \$2.80 as Percent of ATAT Income (4)
1	11.5	6.9	1.6	13.5
2	9.5	10.4	3.4	10.0
3	9.7	10.0	5.8	7.1
4	9.1	11.0	7.5	5.6
5	9.5	10.8	9.3	4.7
6	9.6	10.4	11.0	3.9
7	9.0	10.4	13.2	3.1
8	10.2	9.3	14.8	2.6
9	10.6	10.7	18.5	2.2
10	11.1	10.1	17.0	1.2
Total	100	100	100	3.4

achieved by a change in the minimum wage are dissipated by the tax and transfer system. Under a number of transfer programs, transfers are automatically reduced when wage income increases. In addition, any additional wage income will be subject to taxation by several taxes. To determine the effect of a change in the minimum wage on the net income of households, it is necessary to subtract from the increment in wage income the taxes that will be paid and the reduced transfers that will be received. Elsewhere, we have developed estimates of effective marginal tax rates by household size and income.<sup>5</sup> For most households, marginal tax rates are in the 35–60 percent range. Each household in the sample is assigned a marginal tax rate which is assumed constant for that household when its income varies.

## II. Simulation of Distributional and Efficiency Effects

### A. Increase from \$2.30 to \$2.80; Demand Elasticity Zero

In this section we consider the effects of an increase in the minimum wage from \$2.30 to \$2.80 in 1976. The choice of a \$0.50 increase (an increase of 22 percent) was

largely arbitrary, but reflected a desire to avoid such a small change that the effects would be imperceptible and such a large change that it would be politically unlikely and would directly affect many moderate wage workers.

Several assumptions underlie the results reported in this section. The elasticity of demand for low-wage labor is assumed to be zero; therefore, no disemployment effect would occur in the aggregate. Further, we suppose that no new workers will enter the labor force and replace original workers, and that each of the original workers will continue to work the same number of hours. These conditions imply that the wage income of each affected worker will change in proportion to the change in his wage rate. Downward-sloping labor demand curves will be considered later.

A more difficult question concerns the effect of the change in the minimum on those initially earning less than \$2.30. While those earning between \$2.30 and \$2.80 can reasonably be assumed to have their wages raised to \$2.80, no such obvious treatment exists for the under \$2.30 group. In theory, this group might be affected in many ways, depending in part on whether the reason for the subminimum wage is misreporting, non-compliance, or employment in an uncovered job. In the face of such uncertainty, we have assumed that the subminimum wage rates

<sup>5</sup>The procedures we followed in calculating marginal tax rates are described more fully in our 1981 paper.

TABLE 2—CHANGES IN HOUSEHOLD INCOME WITH NO DISEMPLOYMENT

Decile	No Marginal Tax Rate Adjustment			With Marginal Tax Rate Adjustment			
	After-Tax Income (\$ billion) (1)	Gross Income Added by Minimum Wage Increase (\$ million) (2)	Net Gains of Income Class (\$ million) (3)	Net Gains of Income Class (\$ million) (4)	Percent of Households that Gain (percent) (5)	Distribution of Income Be- fore Increase (percent) (6)	Distribution of Income Af- ter Increase (percent) (7)
1	\$23.5	\$574.0	\$453.8	\$206.2	15.5	2.14	2.16
2	42.6	622.0	402.8	200.6	14.6	3.88	3.89
3	58.1	623.0	323.8	173.7	13.9	5.29	5.31
4	71.9	591.0	221.7	138.0	13.0	6.55	6.56
5	85.1	582.0	143.6	103.8	14.2	7.75	7.76
6	97.8	544.0	41.3	43.6	13.9	8.90	8.91
7	118.0	526.0	-81.0	-29.2	12.4	10.73	10.73
8	137.3	541.0	-171.6	-82.6	13.2	12.49	12.48
9	168.7	525.0	-343.4	-189.6	12.4	15.35	15.33
10	296.0	528.0	-993.7	-516.7	11.2	26.92	26.87
Total	1099.6	5655.0	0	47.8	13.4	100	100
Gini							
Coefficient							
of Overall							
Distribution	.3925	.3903	.3903	.3914	-	.3925	.3914

increase to maintain the same proportion of the minimum wage. In other words, all wages below \$2.30 are assumed to increase by 22 percent, the percentage increase in the minimum wage.

The results of this simulation are reported in Table 2. In order to indicate the significance of marginal tax rates applying to the gains and losses, columns (2) and (3) give the distributional effects *before* this adjustment is made. Column (2) shows the distribution of the gross increase in wage income that results from higher wage rates for low-wage workers. As would be expected from the data in Table 1, the \$5.6 billion total increase is very evenly distributed among income classes, with almost half the total benefit going to the upper half of the income distribution.<sup>6</sup>

Of course, someone must bear the cost of the increase in wage income for low-wage

workers, and the next step is to incorporate these income losses into the analysis. We assume that the \$5.6 billion cost of the increase in the minimum wage falls on all households in proportion to their disposable income.<sup>7</sup> (This would be the approximate effect if the wage rate increase leads to an increase in the price level.) Subtracting the estimated cost for each decile from the gains given in column (2) yields column (3), the net change in income for each income class. On balance, the increase in the minimum wage does redistribute income downward, but the net effect is small. The lowest six deciles gain a total of \$1.6 billion, and this equals the loss to the top four deciles in this case with no disemployment effects.<sup>8</sup> For the lowest decile alone, the net gain is less than 2 percent of its initial disposable income.

<sup>6</sup>We were surprised that the total increase in wage income from a 22 percent increase in the minimum wage was only \$5.6 billion, or 0.5 percent of total disposable income. It should be noted, however, that the total wage income of all workers earning less than \$2.80 per hour was only \$38 billion, so this represents an average increase of 15 percent for the affected workers.

<sup>7</sup>The effects of two alternative methods of allocating the cost (to factor income and to capital income) are reported in our 1981 paper. In general, it appears that the final results are insensitive to the way costs are shifted.

<sup>8</sup>Again, our 1981 paper investigated the effects of a larger minimum wage increase without appreciable differences with the results reported here.

Columns (4) through (7) show the distributional effects when we incorporate the impact of marginal tax rates on the gains and losses. Both the gains and losses that households bear are reduced by the marginal tax rates in the tax and transfer systems. Column (4) gives the net change in income for each income class after this adjustment. Since marginal tax rates are quite high, this has a pronounced effect on the size of gains and losses. Note that now the total net gain to the lowest six deciles is only \$0.9 billion. The net gain for the lowest decile is \$0.2 billion, an increase of less than 1 percent in disposable income.

It is important to recognize that a change in the minimum wage redistributes income within as well as across income classes. Column (5) gives the percentage of households in each income class that gain. Although there is a gain in the aggregate income of the lowest decile, only 15.5 percent of these households actually gain while the remaining 84.5 percent are made worse off. (This is because most low-income households have no low-wage workers, and so they share in the cost but receive no benefits.) An overwhelming majority of low-income households bears a net cost from this policy, while 11 percent of the households in the wealthiest decile gain on balance. It is tempting to infer from this evidence that minimum wages are horizontally inequitable since among those with the same incomes, some gain and some lose. Household income, however, may not be the appropriate index for determining equity in this case.

Finally, columns (6) and (7) show the impact on the percentage shares of income received by each income class. The combined shares of the lowest six deciles rise by a scant 0.07 percentage points, while the increase for the lowest decile is 0.02 percentage points. We also give the related Gini coefficients for the various distributions.

Since the basic assumption of this simulation—no disemployment effects—is quite favorable to finding a significant distributional impact, the small size of the net effects is perhaps the most striking feature of the results.

### *B. Distributional Impact of Disemployment Effects*

Almost all of the quantitative research on minimum wages has centered on determining whether a higher minimum reduces employment of low-wage workers. There seems little doubt that there is some reduction in employment, but the magnitude remains in doubt. In this section, we shall evaluate how the distributional effects of increasing the minimum from \$2.30 to \$2.80 are affected by disemployment effects of various sizes. In addition, we provide estimates of the reduction in national income that occurs when employment falls in response to the higher minimum.

The crucial factor determining the reduction in employment is the elasticity of demand for low-wage labor. We have already considered in detail the case where the elasticity is 0.0; we now turn to the implications of three other elasticities (in absolute value): 0.2, 0.5, and 1.0. Recent estimates of demand elasticities in the literature include Kim Clark and Richard Freeman's estimate of 0.5 and Daniel Hamermesh's estimate of 0.3 in the short run and 0.75 to 1.0 in the long run, so the range of values investigated here seems to span the most reasonable possibilities. For each simulation, it is assumed that the elasticity of demand is the same for every low-wage worker. Since the percentage increase in wage rates differs among workers, this means that the percentage reduction in employment will also vary. Workers with the lowest initial wages bear larger disemployment effects. Recall, however, that we have assumed that all those initially earning less than \$2.30 will receive a 22 percent increase in their wage rates, so employment will fall to the same extent for this entire group.

It should be clear that as long as we are only interested in the overall changes in income of large groups of households, we do not have to specify exactly how the decrease in employment will be allocated among these households. The expected value of the change in earnings for each worker is all that matters to calculate the aggregate income gains or losses for deciles of households.

TABLE 3—DISEMPLOYMENT AND THE DISTRIBUTION OF GAIN  
(\$ million)

Decile	Net Gain of Income Class when Demand Elasticity is			
	0.0	0.2	0.5	1.0
1	\$453.8	\$342.2	\$173.4	\$-108.1
2	402.8	284.7	104.6	-195.4
3	323.8	207.9	29.9	-266.6
4	221.7	114.0	-52.5	-329.9
5	143.6	39.9	-121.4	-390.4
6	41.3	-53.2	-201.4	-448.7
7	-81.0	-168.6	-308.3	-540.9
8	-171.6	-252.7	-394.1	-629.6
9	-343.4	-423.2	-554.6	-773.6
10	-993.7	-1055.4	-1168.4	-1356.8
Total	0.0	-964.4	-2492.8	-5039.9
Net Loss				
Net Gain	1.0	1.98	9.1	-
Gini Coefficient	.3903	.3907	.3914	.3925
Gini Coefficient (with Leisure Adjustment)	.3903	.3905	.3909	.3915

Table 3 shows how the net gain of each income class varies with the elasticity of demand when marginal taxation effects are ignored. As expected, the greater the demand elasticity, the smaller the net gain (or larger the net loss) of each decile. For instance, the net gain by the lowest three deciles is \$1.2 billion when the elasticity is zero, but is \$0.8 billion with an elasticity of 0.2, and is only \$0.3 billion when the elasticity is 0.5. Note also that some income classes which gain when the elasticity is zero actually lose when it is positive. In the zero demand elasticity case, the lowest six deciles gain; when the elasticity is 0.5, only the lowest three gain. With a demand of unit elasticity, all income classes lose, as is to be expected. With a unit elasticity, the wage income of every affected low-wage worker is unchanged, but the rise in the price level reduces the purchasing power of wage (and all other) income so everyone loses.

The reduction in national income is equal to the net gain aggregated across all deciles. The loss in national income varies with the elasticity of demand, but for plausible values this loss appears quite sizable when compared to the net gain of lower income classes. Even when the elasticity is as low as 0.2, the

reduction in national income (\$1 billion) is as large as the entire net gain of the lowest five deciles. Put differently, the net loss of the top five deciles is 1.98 times as great as the net gain for the lowest five deciles. With a demand elasticity of 0.5, the reduction in national income (\$2.5 billion) is eight times as large as the net gain of the lowest three deciles, the only deciles that gain in this case. These estimates suggest that the efficiency cost of increasing the income of lower-income households by raising the minimum wage is quite high.

Marginal tax rates will operate to reduce the gains and losses of households. Table 4 shows the net gain of each income class for each elasticity variant after marginal tax rates have been applied. All gains and losses are reduced sharply. The net gain to the lowest three deciles, for example, is under \$0.2 billion in the 0.5 elasticity case, a gain of only 0.2 percent in the disposable income of these classes. It seems clear that the impact of a 22 percent increase in the minimum wage on the distribution of income is extremely small when marginal taxation and disemployment effects are incorporated.

One way to depict the efficiency-equity tradeoff inherent in these results is to de-

TABLE 4—DISEMPLOYMENT AND NET GAINS AFTER MARGINAL TAXATION  
(\$ million)

Decile	Net Gains of Income Class when Demand Elasticity is			
	0.0	0.2	0.5	1.0
1	\$206.2	\$156.9	\$82.8	\$-40.6
2	200.6	145.9	63.7	-73.1
3	173.7	117.2	32.7	-109.0
4	138.0	79.4	-8.4	-154.9
5	103.8	42.6	-49.4	-202.6
6	43.6	-15.2	-103.4	-250.5
7	-29.2	-86.6	-172.7	-316.5
8	-82.6	-141.2	-229.0	-375.5
9	-189.6	-241.1	-318.3	-447.0
10	-516.7	-556.2	-615.5	-714.3
Total	+47.8	-498.3	-1317.8	-2684.0
Change in				
Tax Revenue	-47.8	-466.1	-1175.0	-2355.9
Gini Coefficient	.3914	.3916	.3919	.3924
Gini Coefficient (adjusted for Leisure)	.3914	.3915	.3916	.3919

termine how egalitarian a social welfare function would have to be to justify using minimum wages to redistribute income. Using a Cobb-Douglas social welfare function,<sup>9</sup> we find that the relative social valuation of dollars of income to the bottom decile relative to the top decile must be approximately 26 times to justify minimum wages when demand elasticity is 0.5. Moreover, this calculation is likely to overstate the attractiveness of minimum wages as social policy because it focuses only on changes in the average income of each decile and ignores redistribution within each decile. If the social welfare function valued negatively horizontal redistribution within an income class (as we would argue it should), it would have to be even more egalitarian with respect to vertical redistribution to show an improvement resulting from an increase in the minimum wage.

<sup>9</sup>In particular, let Social Welfare =  $\prod_{i=1}^{10} Y_i^{\alpha_i}$  where  $i$  indexes deciles,  $Y_i$  is average income in the decile, and  $\alpha_i$  is the decile's weight in the social welfare function. If all the  $\alpha$ 's are equal then the social welfare function values equally the same percentage change in  $Y$  for any decile. Parameterizing the pattern of  $\alpha$ 's to be exponential:  $\alpha_i = \exp(-\beta i)$ , we can find the value of  $\beta$  needed for the minimum wage increase to improve social welfare.

One final adjustment to our results would be to recognize the value of gained leisure when employment falls as a result of a higher minimum wage. Valuing leisure time would clearly reduce the efficiency loss and increase the equity gain of the minimum wage. Michael Hurd develops some estimates of the welfare cost of unemployment. He shows that the welfare cost is highly nonlinear in the duration of unemployment; hence the welfare cost of one person unemployment for six months is much greater than the welfare cost of six persons unemployed for a month each. Although we do not know the average duration of the added unemployment caused by a higher minimum wage, a sensible assumption might be to assume that it is similar to the duration pattern in 1975, a high unemployment year that Hurd studies. In that case, using Hurd's results, the average welfare cost of unemployment is 53 percent of the lost income. It is worth noting that Hurd's method may understate the welfare loss of unemployment by assuming that employed workers can equate their reservation wage with the market wage on the margin; if minimum wages cause rationing of work hours, then Hurd overestimates the opportunity cost of market work.

The last line of Table 4 adjusts the Gini coefficients for the value of leisure time gained by the unemployed using Hurd's method and shows that the equalizing effect of minimum wages is still quite small. Since the total loss in welfare is of course cut by 47 percent, and the Gini coefficient is reduced by the leisure time adjustment, the equity-efficiency tradeoff is improved. However, even in the 0.2 elasticity case, it still requires a billion dollars in social cost (lost income and tax revenue) for each .0010 change in the Gini coefficient as opposed to a .0009 change in the Gini coefficient when leisure is not valued.

### III. Conclusion

When considering both costs and benefits, increasing the minimum wage by 22 percent (and assuming no disemployment) has an equalizing effect on the distribution of household income, but the effect is extremely small. Note that, in the basic simulation (Table 2), the net gain to the lowest decile is only \$0.2 billion from a 22 percent increase in the minimum wage, yet that decile already receives more than 60 times that amount in government transfers. The even distribution of benefits across deciles, the small contribution the earnings of low-wage workers make to total incomes, and the impact of marginal taxation interact to make the minimum wage a very weak redistributive policy.

Increasing the minimum wage redistributes income within income classes as well as across income classes. More than 80 percent of low-income households are harmed by the minimum wage, while more than 10 percent of high-income households actually gain. On the surface, these findings suggest a substantial degree of horizontal inequity.

When disemployment effects are taken into account, the gains to lower-income classes are diminished and the losses to upper-income classes are greater. Even if the elasticity of demand for low-wage labor is as low as 0.2, the reduction in national income is as large as the entire gain to the lower half of

the income distribution when marginal taxation effects are ignored, and the reduction in national income is about twice as large as the net gain to the lower half of the income distribution when they are incorporated. Although valuing leisure mitigates these results somewhat, it still appears that efficiency losses are probably significant, especially when compared to the small changes in the Gini coefficient.

### REFERENCES

- Browning, Edgar K. and Johnson, William R., *The Distribution of the Tax Burden*, Washington: American Enterprise Institute, 1979.
- Clark, Kim and Freeman, Richard, "How Elastic is the Demand for Labor?," Working Paper No. 309, National Bureau of Economic Research, Cambridge, 1979.
- Gramlich, Edward M., "Impact of Minimum Wages on Other Wages, Employment, and Family Incomes," *Brookings Papers on Economic Activity*, 2:1976, 409-51.
- Hamermesh, Daniel, "Subsidies for Jobs in the Private Sector," in John Palmer, ed., *Creating Jobs*, Washington: Brookings Institution, 1978.
- Hurd, Michael, "A Compensating Measure of the Cost of Unemployment to the Unemployed," *Quarterly Journal of Economics*, September 1980, 95, 225-43.
- Johnson, William R. and Browning, Edgar K., "Minimum Wages and the Distribution of Income," in *Report of the Minimum Wage Study Commission Volume VII*, Washington: USGPO, 1981.
- Kniesner, Thomas J., "Low Wage Workers, Who Are They?," in Simon Rottenberg, ed., *The Economics of the Minimum Wage*, Washington: American Enterprise Institute, 1981.
- Welch, Finis, *Minimum Wages*, Washington: American Enterprise Institute, 1978.
- U.S. Congress, Congressional Budget Office, "Poverty Status of Families under Alternative Definitions of Income," Background Paper No. 17, Washington, January, 1977.

# Social Security and Household Saving in an International Cross Section

By ERKKI KOSKELA AND MATTI VIRÉN\*

It is commonly believed that the Social Security system serves to depress the level of household saving in the economy because an individual who expects to receive Social Security benefits will reduce his personal saving. Recent theoretical research has suggested, however, that the relationship is a priori ambiguous; by inducing changes in retirement and/or intergenerational transfers, Social Security can either increase or decrease aggregate household saving. Thus ultimately, the question of whether Social Security reduces saving must be settled by empirical investigation.

Since Martin Feldstein (1974) presented time-series evidence from the United States, according to which the Social Security system has depressed household saving, numerous empirical studies have been done on the subject. Nevertheless, the existing cross-section and time-series evidence from single countries and cross-country evidence is quite mixed (see Feldstein, 1974, 1977; Robert Barro and Glenn MacDonald, 1979; George von Furstenberg, 1979; George Kopits and Padma Gotur, 1980). Differences in data basis, specification of the savings function, and quantification of the Social Security variables are possible reasons for these mixed findings.

This paper provides one further test for the relationship between household saving and Social Security. This is motivated by two considerations which we think have been largely neglected in previous empirical analyses. First, we use the "disequilibrium" sav-

ing hypothesis by Angus Deaton (1977), which has rigorous theoretical foundations and leads up to a relatively simple specification of the savings function. Moreover, there is some international evidence, which is favorable for using this hypothesis as a general frame of reference when explaining the behavior of savings ratios over time (see our referenced papers). Second, we use a large international data sample from sixteen OECD countries over the period 1960-77. To guarantee a reasonably good quality of data, only OECD countries have been included into our data sample. Compared with earlier cross-country studies, the data is up to date and includes the "volatile" years of the mid-1970's. What is more important, it is based in majority of cases on the current System of National Accounts (SNA), so that household saving to be explained does exclude compulsory insurance saving which consisted of a considerable part of the household saving in the former SNA (see Appendix).

## I. Empirical Results

We turn now to consider empirical results. The savings function to be estimated is of the form

$$(1) \quad (s/y)_t = b_1 q_t + b_2 g_t + b_3 (s/y)_{t-1} \\ + b_4 r_t + b_5 (\Delta U)_t + b_6 SS_t + b_7 OLD_t \\ + b_8 PR_t + \sum_{j=1}^{16} d_j D_j + u_t.$$

The first three terms on the right-hand side represent the basic specification of the disequilibrium saving hypothesis, according to which the savings ratio depends on real income rate of change and inflation rate "innovations,"  $g_t$  and  $q_t$ , and the lagged savings

\*Professor of economics, University of Helsinki, and doctor of economics, Research Institute of the Finnish Economy, respectively. We are indebted to Timo Rajakangas for research assistance and an anonymous referee for helpful comments. Financial support from the Economic Research Council of Nordic Countries and from the Co-Operative Banks' Research Foundation is gratefully acknowledged.

ratio,  $(s/y)_{t-1}$ . Innovation terms in equation (1) have been derived both in the case of *constant* and *static expectations hypotheses* for rate of change of real income and inflation rate. Thus in the former case,  $g_t = \Delta \log y_t$  and  $q_t = \Delta \log p_t$ , while in the latter case,  $g_t = \Delta \Delta \log y_t$  and  $q_t = \Delta \Delta \log p_t$ , where  $y_t$  is the real (households' disposable) income and  $p_t$  the implicit deflator of consumption expenditure.<sup>1</sup>

In equation (1),  $r_t$  indicates the real rate of interest (which is simply the nominal rate of interest  $R_t$  in the case of constant expectations, while in the case of static expectations it is the nominal rate of interest minus the lagged inflation rate). The variable  $(\Delta U)_t$  stands for the difference in the unemployment rate. This term can be interpreted as a proxy for real income uncertainty, which affects household saving positively according to the 'uncertainty' hypothesis.

The terms  $SS_t$ ,  $OLD_t$ , and  $PR_t$  represent the Social Security variables in the present analysis. The first,  $SS_t$ , denotes the Social Security benefits relative to the population over 65 divided by per capita *GDP*. In this context, two benefit measures were used: a broad one which includes all Social Security benefits and a narrow one which includes only the benefits received by the aged, survivors, and disabled. The latter concept was used, for example, by Barro and MacDonald (1979) and Feldstein (1977, 1980). We also experimented with some cruder measures of Social Security, such as Social Security benefits divided by *GDP*,  $SB_t$ . In what follows, only the estimation results based on the narrow measure of Social Security benefits  $SS_t$  are reported; *results, however, were practically identical with all three measures of Social Security benefits*. The second term,  $OLD_t$ , is the ratio of population

over 65 to total population, and the third,  $PR_t$ , is the participation rate of population over 65. Finally,  $u_t$  is the error term (see the Appendix for a detailed description of data and data sources).

A major problem in studying effects of the "pay-as-you-go" Social Security system is how to estimate the level of benefits individuals anticipate receiving when they retire. From this point of view, our measure(s) of Social Security,  $SS_t$ , is a rather crude one. "More sophisticated" quantifications of expected Social Security benefits or Social Security wealth do not solve this problem either (see Feldstein, 1974, and von Furstenberg). Therefore, we have abstained from extensive quantification attempts of benefit expectations.<sup>2</sup>

As pointed out above, a demographic composition variable,  $OLD_t$ , and a participation rate variable,  $PR_t$ , are also included in equation (1). The latter might capture the possible induced retirement effect of Social Security on household saving, which in turn might offset the asset substitution effect. In order to account for possible simultaneity between  $(s/y)_t$  and  $PR_t$ , an equation for  $PR_t$  was specified and (1) was estimated by two-stage least squares (2SLS). The specification adopted for participation rate was of the form:

$$(2) \quad PR_t = a_1 OLD_t + a_2 SS_t + a_3 U_{t-1} + a_4 PR_{t-1} + \sum_{j=1}^{16} c_j D_j,$$

where  $U_{t-1}$  is the lagged unemployment rate term capturing possible discouragement effects on the participation rate.

Both equations (1) and (2) include individual intercepts for each country. These variables account for institutional and other related differences in the household saving ratio and participation rate across countries (the former is an important, but mostly neglected area of research; see, however, Franco

<sup>1</sup> Obviously these expectations hypotheses, which represent two extreme alternatives from the point of view of the *adaptive expectations hypothesis*, are not very sophisticated. The number of observations for each country, however, does not make it possible to use time-series models for "innovation" terms. Moreover, there is some international evidence, according to which the disequilibrium savings function specification with constant expectations is not unjustified even for the data of 1970's (see our referenced papers).

<sup>2</sup> Moreover, as has been pointed out by Jeffrey Carmichael (1982), Social Security wealth variables seem to have no strong theoretical justification.

TABLE 1—ESTIMATES OF THE SAVINGS RATE AND PARTICIPATION RATE EQUATIONS

	[1]: $s/y$	[2]: $s/y$	[3]: $PR$	[4]: $s/y$
$q_t$	.1654 ( 5.27)	.2044 ( 4.60)		.1634 ( 5.49)
$g_t$	.3811 (11.73)	.2766 ( 8.46)		.3868 (12.96)
$s/y_{t-1}$	.5869 (13.11)	.7556 (13.70)		.5854 (14.36)
$r_t$	.1108 ( 1.34)	-.0480 ( 1.10)		.0947 ( 1.26)
$(\Delta U)_t$	.0043 ( 3.43)	.0030 ( 1.88)		.0042 ( 3.61)
$SS_t$	-.0043 ( 0.48)	-.0080 ( 0.81)	-.0317 (16.19)	.0018 ( 0.66)
$OLD_t$	-.2304 ( 1.19)	-.2263 ( 1.04)	-.1779 ( 1.90)	-.2950 ( 1.55)
$PR_t$	-.0335 ( 0.53)	-.1006 ( 1.31)		-.0504 ( 0.76)
$PR_{t-1}$			.7679 (26.97)	
$U_{t-1}$			-.0022 ( 0.04)	
Individual Intercepts:				
Australia	.0402 ( 1.76)	.0657 ( 2.46)	.0588 ( 5.58)	.0448 ( 1.97)
Austria	.0404 ( 1.38)	.0656 ( 1.92)	.0592 ( 4.28)	.0467 ( 1.61)
Belgium	.0660 ( 2.43)	.0784 ( 2.47)	.0468 ( 3.62)	.0725 ( 2.70)
Canada	.0235 ( 1.02)	.0585 ( 2.16)	.0610 ( 5.55)	.0284 ( 1.23)
Finland	.0058 ( 0.25)	.0466 ( 1.76)	.0606 ( 5.87)	.0092 ( 0.41)
France	.0526 ( 1.76)	.0770 ( 2.20)	.0598 ( 4.15)	.0601 ( 2.01)
Germany	.0709 ( 2.35)	.0839 ( 2.38)	.0677 ( 4.90)	.0762 ( 2.58)
Greece	.0652 ( 2.12)	.0944 ( 2.62)	.0809 ( 5.82)	.0705 ( 2.33)
Italy	.0858 ( 3.27)	.0940 ( 3.05)	.0633 ( 5.56)	.0886 ( 3.55)
Japan	.0758 ( 2.25)	.1091 ( 2.68)	.0920 ( 5.80)	.0847 ( 2.46)
Netherlands	.0533 ( 2.20)	.0732 ( 2.62)	.0636 ( 6.14)	.0555 ( 2.42)
Portugal	.0428 ( 1.35)	.0816 ( 2.18)	.0785 ( 5.18)	.0514 ( 1.59)
Sweden	.0277 ( 0.90)	.0585 ( 1.65)	.0691 ( 4.79)	.0343 ( 1.12)
Switzerland	.0670 ( 2.06)	.0845 ( 2.24)	.0826 ( 5.53)	.0738 ( 2.28)
U.K.	.0323 ( 1.09)	.0620 ( 1.77)	.0720 ( 5.12)	.0394 ( 1.32)
U.S.	.0333 ( 1.23)	.0612 ( 1.93)	.0720 ( 5.67)	.0386 ( 1.44)
$N$	240	224	240	240
$R^2$	.9623	.9536	.9918	.9643
$SEE$	.0113	.0127	.0060	.0107
$\chi^2_3$	1.3898	7.3693	54.7059	1.2941

Notes: The  $t$ -statistics are shown in parentheses;  $\chi^2_3$  stands for a test statistic for first-order autocorrelation computed from a contingency table. The corresponding critical value at 5 percent level is 7.81.

Modigliani, 1970). Moreover, the individual country intercepts in (1) can also contain differences in the expected levels of real income rate of change and inflation rate across countries (see Deaton, 1977).

The estimation results are presented in Table 1, where four sets of estimates have been tabulated: [1] contains the ordinary least squares (*OLS*) estimates of the savings function in the case of constant expectations; [2], the corresponding estimates in the case of static expectations; [3] reports the *OLS* estimates of the participation rate equation, and finally, the *2SLS* estimates of the savings function with constant expectations are presented in [4].<sup>3</sup>

<sup>3</sup>Estimations in Table 1 are unweighted ones. Weighting the country observations by population gave practically identical results.

The disequilibrium savings function hypothesis fits the data very well: there seems to be no first-order autocorrelation, all coefficient estimates are of expected sign, reasonable magnitude, and highly significant.<sup>4</sup> According to the estimates, unanticipated inflation, as well as unanticipated real income growth, affect the savings ratio positively. As far as variables  $r_t$  and  $(\Delta U)_t$ , in addition to the disequilibrium saving hypothesis, are concerned, the real rate of interest does not show up very well in contrast to the change in unemployment, which seems to increase

<sup>4</sup>High fit is also obtained by estimating the savings function (1) for time-series data of each country. The  $R^2$  varies between .74 and .98 with the median being .93. In the case of Portugal only, we could not reject the hypothesis that all coefficients are zero (by using  $F$ -test with 5 percent significance level).

the fit in all specifications [1], [2], and [4]. Hence, the real income uncertainty hypothesis proxied by  $(\Delta U)_t$  does not contradict our data.

The participation rate equation [3] fits the data very well, all coefficient estimates being of expected sign and, with the exception of the unemployment rate, rather precisely estimated. And most interestingly, a rise in Social Security benefits tends to induce earlier retirement.<sup>5</sup>

As far as the performance of the Social Security variables in the savings rate equations are concerned, it is easy to summarize their role: *the Social Security variables, used in this study, have no effect on the household saving ratio.* This result is robust to changes in the definition of  $SS_t$  variable, to changes in the combination of variables (including  $r_t$ ,  $(\Delta U)_t$ ,  $OLD_t$ , and  $PR_t$ ) to the procedure of weighting country observations, to the estimation method and to the expectations hypotheses. Our results are consistent with findings by Barro and MacDonald, on the one hand, and by Kopits and Gotur, on the other hand, but do not lend support to the "Social Security depresses saving" proposition forcefully advocated by Feldstein (1974, 1977, 1980).

One may wonder what causes this huge difference between our results and those of Feldstein (1977, 1980). Feldstein uses the private sector savings rate as the dependent variable and his data sample and time period differ from those used in this study. Moreover, cross-section data with fifteen and twelve countries are used respectively in Feldstein (1977, 1980). All differences in the savings ratios across countries will thus be attributed to growth rate, demographic and Social Security variables so that in pooled cross-country time-series data, the individual country intercepts are restricted to be the same (the different treatment of individual country intercept may matter as is evident from Barro and MacDonald).

This procedure of restricting individual country intercepts to be the same is not

appropriate in our sample: the hypothesis that intercepts are equal can be rejected. The value of the  $F$ -test for equation [1] is  $F_{15,233} = 5.5071$ , which is significant at all standard significance levels.

But even though we impose the equality restriction for country intercepts, results are not affected. Still, Social Security variables have no effect on household saving; the coefficient estimate of  $SS_t$  is  $-.0006$  ( $t$ -statistic  $= -.18$ ). Moreover, if one compiles the cross-section data for a similar set of countries as in Feldstein (1980), and uses the same variables except the  $SS_t$  variables instead of Feldstein's "new retiree replacement ratio," then the "social security depresses saving" proposition gets no support.<sup>6</sup> Feldstein's results are not robust enough to allow for his strong conclusions about the depressing effect of Social Security on household saving.

There are at least three ways of explaining the zero effect of Social Security on household saving: (i) an extended life cycle view with endogenous labor supply: the "asset substitution" and "induced retirement" effect of Social Security cancel each other; (ii) a dynastic cross-generation view: the benefi-

<sup>6</sup>The following savings function, which is similar to equation (1) in Feldstein (1980), was estimated with a cross-section data over the period 1971–78:

$$(s/y)_t = b_0 + b_1 G_t + b_2 AGE_t + b_3 DEP_t + b_4 SS_t + b_5 PR_t + u_t,$$

where  $s/y$  is the private savings rate,  $G$  is the growth rate of total private income,  $AGE$  is the ratio of the number of retirees over age 65 to the population aged 20–65,  $DEP$  is the ratio of the number of younger dependents to the working age population, and  $SS$  and  $PR$  are the Social Security and participation rate variables, as defined earlier. On the basis of  $OLS$  estimates with data samples of 10 and 15 countries, respectively, we were unable to reject the hypothesis that all coefficients are equal to zero; values of the  $F$ -test were as low as .51 and .38 (Feldstein's sample consisted of 12 countries and  $G$  was computed over the period 1960–75, see Feldstein, 1980, p. 234). The whole evidence in Feldstein (1980) does seem to rest on his "new retiree replacement ratio" variable,  $B/E$ , and then only when the observations are weighted by population! If, in turn, one uses the replacement ratio variable which is used in this study, in Feldstein (1977), and in Barro and MacDonald (1979), no evidence is found for Feldstein's conclusion. The set of results (both with Feldstein's data sample and with our data samples) is available upon request from the authors.

<sup>5</sup>There seems, however, to be first-order autocorrelation, possibly biasing the test results, according to the  $\chi^2_2$ -statistics. This is not surprising, given the small annual variation of the participation rates.

ciaries of Social Security will transfer their gains to the generations that will have to pay more taxes as a result of a rise in Social Security benefits so that there is no change in aggregate savings; (iii) a liquidity view: a rise in Social Security benefits financed by taxes decreases consumption of economically active people by the full amount of taxes, and if pensioners were also liquidity-constrained, then they would consume all their benefits. Thus there would be no change in household saving (see Walter Dolde and James Tobin, 1980).

While it lies beyond the scope of this note to try to distinguish between these explanations, we may point out that the estimates with respect to the  $PR_t$  variable give some support to the induced retirement effect, even though the evidence does not appear to be very strong.

#### APPENDIX

$C_t$ : Private final consumption expenditure from *National Accounts of OECD Countries*, various issues. The former system of SNA is used for Belgium, Greece, Portugal and Switzerland, the present SNA for other countries. Swedish data is unpublished, and obtained from Swedish National Central Bureau of Statistics.

$S_t$ : Household saving from *National Accounts of OECD Countries*, various issues. Figures include the nonprofit institutions serving households with all countries except Canada and Italy.

$Y_t$ : Households' disposable income from *National Accounts of OECD Countries*, various issues.

$P_t$ : The implicit price deflator for private final consumption expenditure.

$SB_t$ : Social Security benefits as percentage of gross domestic product from *The Cost of Social Security*, International Labor Office, various issues.

$SS_t$ : Social Security benefits (total amount of benefits, and the benefits received by the aged, survivors, and disabled) relative to population over 65 divided by per capita gross domestic product from *The Cost of Social Security*, International Labor Office, various issues, and *OECD Labor Force Statistics*, various issues.

$OLD_t$ : The ratio of population over 65 to the total population from *OECD Labor Force Statistics*, various issues.

$PR_t$ : Economically active population over 65 from *Yearbook of Labor Statistics*, International Labor Office, various issues. Missing observations have been extrapolated by the present authors.

$U_t$ : Unemployment rate from *OECD Labor Force Statistics*, various issues.

$R_t$ : Government bond yield from *International Financial Statistics*, various issues.

The data covers the period 1964–77 for Austria and Greece, 1961–77 for France, 1962–77 for Italy and Switzerland, 1970–77 for Japan, 1965–76 for Portugal, and 1960–77 for all other countries.

#### REFERENCES

- Barro, Robert J. and MacDonald, Glenn M., "Social Security and Consumer Spending in an International Cross Section," *Journal of Public Economics*, June 1979, 11, 275–89.
- Carmichael, Jeffrey, "On Barro's Theorem of Debt Neutrality: The Irrelevance of Net Wealth," *American Economic Review*, March 1982, 72, 202–13.
- Deaton, Angus, "Involuntary Saving through Unanticipated Inflation," *American Economic Review*, December 1977, 67, 899–910.
- Dolde, Walter and Tobin, James, "Mandatory Retirement Saving and Capital Formation," unpublished paper, 1980.
- Feldstein, Martin S., "Social Security, Induced Retirement and Aggregate Capital Accumulation," *Journal of Political Economy*, September/October 1974, 82, 905–26.
- , "Social Security and Private Savings: International Evidence in an Extended Life-Cycle Model," in his and Robert P. Inman, eds., *The Economics of Public Services*, London: MacMillan, 1977, 174–205.
- , "International Differences in Social Security and Saving," *Journal of Public Economics*, October 1980, 14, 225–44.
- Kopits, George and Gotur, Padma, "The Influence of Social Security on Household Savings: A Cross Country Investigation,"

*IMF Staff Papers*, March 1980, 27, 161-90.  
Koskela, Erkki and Virén, Matti, "Inflation and Saving: Testing Deaton's Hypothesis," *Applied Economics*, forthcoming.

\_\_\_\_\_ and \_\_\_\_\_, "Saving and Inflation: Some International Evidence," *Economics Letters*, 1982, 9, 337-44.

Modigliani, Franco, "The Life-Cycle Hypothe-

sis and Inter-Country Differences in the Saving Ratio," in W. Eltis, et al., eds., *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, Oxford: Clarendon Press, 1970, 197-225.

von Furstenberg, George, *Social Security versus Private Saving*, Cambridge: Ballinger, 1979.

# Price Controls in a Posted Offer Market

By DON L. COURSEY AND VERNON L. SMITH\*

The effects of price controls have been examined in markets organized under the double auction trading institution. R. Mark Isaac and Charles Plott (1981) and Smith and Arlington Williams (1981) report the following three principal conclusions from twenty-eight double auction experiments:

1) The hypothesis is rejected that non-binding price ceilings (floors) will serve as a focal point in the sense that buyers and sellers are attracted to prices at the ceiling (floor).<sup>1</sup>

2) Evidence is presented to support the hypothesis that nonbinding ceilings (floors) near the competitive equilibrium (*CE*) price will lower (raise) the convergence path of contract prices to the *CE* price. Also, supported is the hypothesis that contract prices tend to converge to the level of a binding price ceiling (floor) from below (above) the control price.

3) Strong empirical evidence was presented which showed that upon the removal of either a binding or a nonbinding price ceiling (floor) prices tended to jump discontinuously above (below) the *CE* level, only returning to that level after a period of adjustment.

This study seeks to determine whether these three conclusions are unique to the double auction institution of exchange by examining the effects of price ceilings in the posted offer institution of exchange. A posted offer market is characterized by sellers who publicly post nonnegotiable prices (that are selected privately) for a commodity at the beginning of each period of trading. Buyers then respond (in random order) by purchasing desired quantities of the commodity from

the seller of their choice. Most retail markets use posted offer pricing, and we ordinarily associate price controls (as in the Nixon Administration) with such markets.

We report below the design and results of thirteen controlled laboratory posted offer experiments using an experimental design consisting of fixed supply and demand schedules. Across independent experiments the level of a price ceiling is varied in 5 cent steps from above to below the *CE* level. In half of the experiments, the price control is removed midway through the trading periods of an experiment ("controlled/non-controlled" or *C/N* experiments), and in other experiments, the price control is imposed upon the market only after the midpoint is reached ("noncontrolled/controlled" or *N/C* experiments). By considering both the *C/N* and *N/C* treatments, more insight into the reported double auction price explosion effect is obtained.

## I. Experimental Design

All of the experiments use the PLATO computer version of the posted offer exchange mechanism as described in Jon Ketcham, Smith, and Williams (1980). Using this system, it is possible to induce value (see Smith, 1976) on the actions of buyers and sellers which define supply and demand schedules that are known to the experimenter, but not to the agents. Our design is characterized by four buyers capable of purchasing three units each and four sellers capable of selling three units each. The resale and cost values of these units are charted on the left in Figure 1.

All of the subjects who participated in the experiment were undergraduate students at the University of Arizona, and all had prior experience in at least one similar posted offer experiment *without* price controls and with different supply and demand parameters. The subjects received a special announcement that

\*University of Arkansas and University of Arizona, respectively. We are grateful to the National Science Foundation for Research support.

<sup>1</sup>A nonbinding price ceiling is a ceiling equal to or greater than the competitive equilibrium price. A binding ceiling is one which is below the competitive equilibrium price.

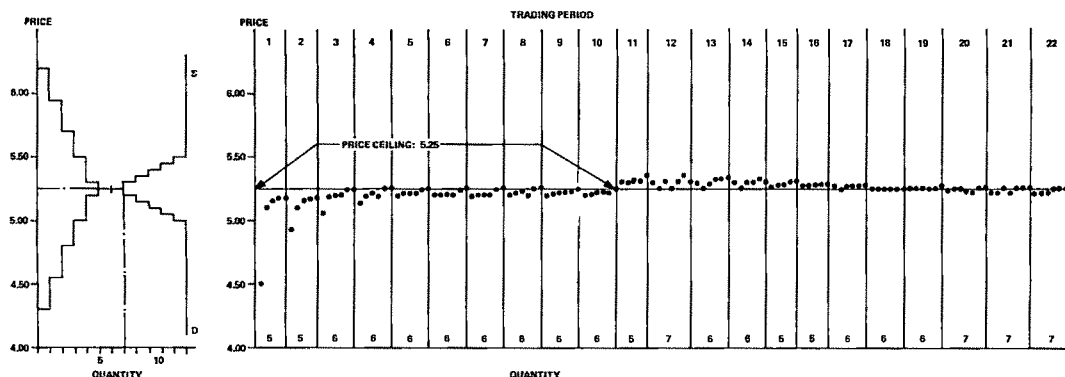


FIGURE 1

price controls would be in effect when appropriate, during each period of trading. Attempted violation of the price ceiling resulted in the rejection of a seller's offer until the price constraint condition was satisfied. Each trading period began with the sellers selecting (privately) the price at which they wished to sell units, and the number of units they wished to sell at this price. Once all of the sellers had finalized their offers, each seller was notified of the other sellers' offer prices, and the buyers were randomly queued to start the buying procedure. When there were no units available from a particular seller, an "out-of-stock" message was displayed. To provide incentive for the marginal units to trade each contract yielded a commission of 10 cents to both the trading buyer and seller. After each buyer had completed this purchase mode, the trading period was over.

All thirteen experiments consisted of twenty trading periods separated into two groups of ten periods each. For the *C/N* type experiments, the price control was in effect for the first ten periods, and not in effect for the second ten periods. Conversely, the *N/C* experiments began with ten periods of noncontrol and ended with ten periods of price control. The ceilings varied in 5 cent steps across independent experiments. Experiments were conducted with controls 5 cents above the competitive equilibrium (experiments P39, P56, P57, and P60), at the competitive equilibrium (P65, P74, and P101), 5 cents below (P85, P88, and P100), and 10 cents below (P81, P84, and P99).

## II. Experimental Results

Previous experimental studies of posted offer (bid) pricing (Fred Williams, 1973; Plott and Smith, 1978) have indicated that prices in this institution tend to converge from above (below) the *CE* price to a level somewhat higher (lower) than the *CE* price. The confirmed tendency of posted offer prices to converge from above strongly suggests that price ceilings at or 5 cents above the *CE* price, and certainly for binding ceilings below the *CE* price, will yield many if not all observed contracts at the ceiling price. In particular it seems quite likely, at least for the first few trading periods, that the focal point hypothesis will be confirmed as sellers, desiring to begin with "high" posted prices focus and "lock on" to the ceiling price even if it is nonbinding. These were our *a priori* expectations. Our most important finding is that these conjectures proved to be false.

The individual results for a typical experiment (P65) are displayed in Figure 1. In Figure 1, all contract prices are plotted in the order in which the contracts occurred. The *CE* price is indicated by the solid line at \$5.25. The periods for which the price ceiling treatment condition was in effect are indicated on each figure.

Overall, we find qualitative continuity between the previous double auction results and our posted offer results. If we examine the convergence behavior of prices over time, the predictions of competitive price theory, with or without ceilings, is strongly sup-

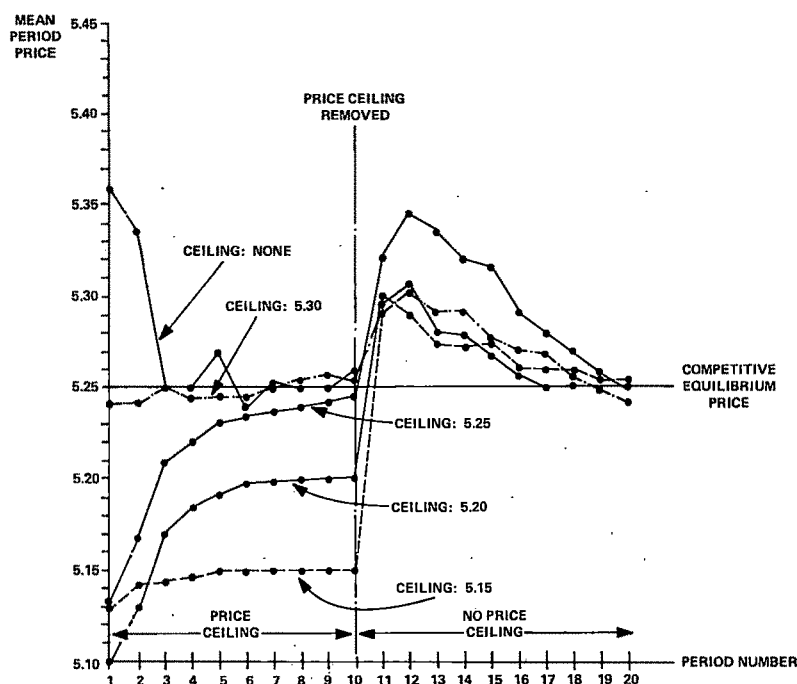


FIGURE 2

ported. This is shown in Figure 2 which plots the time-series of mean price across experiments for each trading period. The graph of the data for the first ten periods, in those experiments with no price ceiling, exhibits the familiar tendency to converge from above as reported in all previous posted offer experiments. This pattern is reversed when there is a price ceiling at any level from 5 cents above to 10 cents below the *CE* price. Hence, the dynamic effect of price ceilings in posted offer markets is qualitatively the same as in double auction markets. In both markets the effect of a ceiling is to cause convergence from below to the constrained or unconstrained competitive price. However, this downward shift in the price convergence path is much more pronounced in double auction markets.

Since sellers in an unconstrained posted offer market tend to post initial prices that are high relative to the *CE* price, why do they tend not to begin with posted prices at the ceiling price, when such ceilings are in effect? The answer appears to be as follows:

Sellers do not know what constitutes a "high" price, that is, they do not "know" the *CE* price, nor do they know whether the ceiling price is binding or not, even if we assumed that the sellers would understand what is meant by a "*CE* price" or a "binding" ceiling. However, from their experience with previous posted offer markets, each seller has learned that if he/she has a higher price than one or more other sellers, this increases the chance that no sales or profit will be made. Hence, an initial impulse to quote prices at the ceiling price is likely to be tempered by the thought that "others may charge less, so that I may fail to make a sale. Perhaps I should begin at a lower price." This type of behavior is certainly consistent with agent experience, and is also consistent with what we observe in markets with a price ceiling.

Price ceilings were removed in nine of the thirteen experiments. The effect on mean price by period is also shown in Figure 2. The removal of either a binding or a non-binding price ceiling produces the jump discontinuity observed in the double auction

experiments. However, unlike the double auction results, we cannot clearly identify a positive relationship between the size of the initial jump and the degree to which the original ceiling was binding. Also, the increase in prices is less explosive than in double auction markets. After the initial increase in prices, following the lifting of a ceiling price, the market tends to converge toward the free market *CE* price. A jump in prices following the removal of price control is often observed in the field, as, for example, with the removal of gasoline price controls in 1980. This behavior is often described as the result of "pent-up" or accumulated unsatisfied demand. But this explanation cannot describe our experimental market behavior since all purchases are for current period demand—there can be no accumulation of demand across trading periods. The phenomena would appear to be explainable only in terms of seller expectations that sales are feasible, and perhaps sustainable, at prices higher than the removed ceiling. If sellers differ in these expectations, posted offer prices are likely to exhibit an increased mean and variance for a few trading periods as sellers grope for a new sustainable level of prices and profits.

### III. Summary

In summary, we found that the standard competitive model is supported with respect to its performance in describing the *equilibrium* tendencies under price control in posted offer markets. Similar results had been reported previously for double auction markets. We also found that certain dynamic "irregularities" in double auction markets also occur in posted offer markets. The de-

pression of initial prices in price ceiling markets has not been anticipated by traditional microeconomic theory. Binding as well as nonbinding price ceilings affect the negotiation process in both trading institutions, and tend to lower initial contract prices. Finally, an initial increase in contract prices follows the removal of either a binding or a nonbinding price ceiling, but ultimately prices approach the *CE* price.

### REFERENCES

- Issac, R. Mark and Plott, Charles R., "Price Controls and the Behavior of Auction Markets: An Experimental Examination," *American Economic Review*, June 1981, 71, 449-59.
- Ketcham, Jon, Smith, Vernon L. and Williams, Arlington W., "The Behavior of Posted Offer Pricing Institutions," Southern Economic Association Meetings, November 5-7, 1980.
- Plott, Charles R. and Smith, Vernon L., "An Experimental Examination of Two Exchange Institutions," *Review of Economic Studies*, February 1978, 45, 133-53.
- Smith, Vernon L., "Experimental Economics: Induced Value Theory," *American Economic Review Proceedings*, May 1976, 66, 274-79.
- \_\_\_\_\_ and Williams, Arlington W., "On Non-binding Price Controls in a Competitive Market," *American Economic Review*, June 1981, 71, 467-74.
- Williams, Fred, "Effect of Market Organization on Competitive Equilibrium: The Multi-Unit Case," *Review of Economic Studies*, January 1973, 40, 97-113.

# Government Debt in an Overlapping-Generations Model with Bequests and Gifts

By JOHN B. BURBIDGE\*

In 1974, Robert Barro argued that if individuals in successive generations were linked by bequests, changes in the stock of government debt or in Social Security programs would have no effect on the steady-state capital stock. Barro assumed that individuals held static expectations and that the size of the population was constant. Martin Feldstein (1974, 1977) had concluded that Social Security would reduce the capital-labor ratio in models that did not admit a bequest motive, and, in his 1976 comment on Barro, argued that the introduction of government debt or Social Security into a perfect foresight, dynamic growth model with bequests would still reduce the capital-labor ratio. Barro replied that Feldstein's criticisms were invalid if the steady-state capital-labor ratio were smaller than the Golden Rule level, but would be correct if the opposite were true. Barro did suggest, however, that the latter possibility, with overaccumulation of capital, might be ruled out in his model, as it was in Miguel Sidrauski (1967), but he was unable to demonstrate that the required behavior would be consistent with utility maximization by finite-lived individuals (1976, p. 345). Although recent work by Willem Buiter (1979), Jeffrey Carmichael (1979, 1982), Truman Bewley (1981a, b), and others has helped to elucidate the nature of overlapping-generations models with bequests and gifts, and the issues in the Barro-Feldstein debate, they remain far from clear. The purpose of this paper is to shed a little more light on the subject.

In what follows, I contend that some of the authors cited above either have specified the individual's optimization problem in an

asymmetric way, or have failed to impose necessary conditions for a sensible optimization problem, or both. There is a natural specification of the "Barro model"<sup>1</sup> that emphasizes the similarity between this model and optimal growth models (see, for example, David Cass, 1965; Peter Diamond, 1973). Here, in steady-state equilibrium, the (after-tax) interest rate,  $r$ , must equal the rate at which individuals discount the utility of their heirs,  $\rho$  ( $r = \rho$  has become known as the "modified Golden Rule") and a meaningful individual optimization problem requires that  $\rho$  exceed  $n$ , the natural growth rate of the economy. My results confirm Barro's hunch about his model; in a Barro model,  $r$  must exceed  $n$ . Much of the Barro-Feldstein 1976 interchange, which assumed that  $r$  could have any relation to  $n$ , is wrong-headed. Barro and Feldstein reach different conclusions about the effects of government debt and Social Security because they assume different individual utility functions. Also, my results contrast sharply with those of Buiter and Carmichael. For example, Carmichael concluded that with intergenerational transfers from parents to children (bequests),  $r$  must exceed  $n$ , but with transfers in the opposite direction,  $n$  must exceed  $r$  (1982, pp. 205-06). Indeed, I believe that the Buiter-Carmichael representation of the Barro model is logically faulty (see Section II).

In Section I, I outline an overlapping-generations model with gifts and bequests, and derive its steady-state properties. In the next section, I discuss the short- and long-run effects of introducing government debt into the model and contrast my results with those

\*Associate professor of economics, McMaster University. I thank Jeff Carmichael, Peter Diamond, Jon Eaton, Brian Ferguson, Alan Harrison, and Bill Scarth for helpful comments on earlier drafts.

<sup>1</sup>For the most part, I use this term in place of the more cumbersome "overlapping-generations model, with gifts and bequests and intergenerationally-dependent preferences," even though I adopt a specification of the model that differs from Barro's.

of others. In Section III, I allow for a labor-leisure choice. It turns out that the second-order conditions guarantee that, with government debt and taxes on first-period work, individuals choose to work less and therefore debt reduces both the capital stock and the level of steady-state utility. My conclusions are summarized in Section IV.

### I. The Barro Model

Barro (1974, p. 1100) wrote the utility function for someone who cares about his heirs as

$$(1) \quad v_t = f_t(c_t^1, c_t^2, v_{t+1}),$$

where  $v_t, c_t^1, c_t^2$  are utility and consumption when young and old for generation  $t$ , and  $v_{t+1}$  is the attainable utility level of the individual's descendant. Buiter (1979, p. 418) and Carmichael (1982, p. 205) have observed that there must be some discounting of  $v_{t+1}$  for  $v_t$  to be bounded, that is, for there to exist a sensible optimization problem. They write (1) as being additively separable in the individual's own consumption and his descendant's utility level:

$$(2) \quad v_t = u(c_t^1, c_t^2) + v_{t+1}/(1 + \rho), \quad \rho > 0,$$

where  $\rho$  is the discount rate applied to the utility level of the individual's immediate successors. Unlike Barro, however, Buiter and Carmichael assume population grows at  $n$ , so that a person has not one heir, but  $1 + n$  heirs. While they build population growth into the constraints that impinge on the individual, they ignore the structure that population growth imposes on the utility function. It is appropriate to write the utility function of someone born at time  $t$  as

$$(3) \quad v_t = u(c_t^1, c_t^2) + \frac{1+n}{1+\rho} v_{t+1} \\ = \sum_{i=0}^{\infty} \left( \frac{1+n}{1+\rho} \right)^i u(c_{t+i}^1, c_{t+i}^2).$$

The utility function in (3) is well-defined

only if the sum converges and therefore  $\rho > n$  is a necessary condition for sensible results.<sup>2,3</sup>

When Buiter (1979, p. 421) and Carmichael (1982, p. 205) permit the individual to care about his or her parents to analyze gifts, they introduce a second problem by assuming that children *discount* the utility of their parents at  $\rho$ . Consistency of the family's consumption plan requires, however, that if the utility of heirs is discounted at  $\rho$  the utility of parents is "reverse-discounted" at  $\rho$ ,<sup>4</sup> that is,

$$(3') \quad v_t = u(c_t^1, c_t^2) + \left( \frac{1+n}{1+\rho} \right)^{-1} v_{t-1} \\ = u(c_t^1, c_t^2) + \left( \frac{1+n}{1+\rho} \right)^{-1} u(c_{t-1}^1, c_{t-1}^2) \\ + \left( \frac{1+n}{1+\rho} \right)^{-2} v_{t-2}.$$

Since someone born at time  $t$  can affect only the utility level of his or her parents,  $v_{t-2}$  is exogenous at time  $t$ . A formulation of the utility function that permits analysis of gifts as well as bequests is equation (4), in which  $\rho$

<sup>2</sup>Without additive separability the restriction would be  $\partial v_t / \partial v_{t+1} < 1$ .

<sup>3</sup>With technical progress at rate  $m$ ,  $c_t^1$  and  $c_t^2$  would be growing at  $m$  in a steady state. If the utility function were homogeneous of degree one, this would mean that  $1 + \rho$  would have to exceed  $(1 + n)(1 + m)$ . I show below that the steady-state interest rate,  $r$ , must equal  $\rho$ , so that the Barro model could never have more than the Golden Rule capital-labor ratio in a steady-state equilibrium. For homogeneous utility functions, only if the degree of homogeneity is less than unity could this result be reversed.

<sup>4</sup>Readers of earlier drafts of this paper have been bothered by this assertion. Drawing on equations (6) and (7) below, one can see why consistency does require reverse discounting. Suppose individuals discount the utility of their heirs at  $\rho_+$  and that of their parents at  $\rho_-$ . Given his or her budget constraint, a person of generation  $t$  contemplating gifts to his parents would equate  $(\partial u / \partial c_{t-1}^1) / (\partial u / \partial c_t^1)$  to  $(1 + r_t)(1 + \rho_-)$ . A person of generation  $t - 1$  contemplating bequests to his children would equate the same ratio to  $(1 + r_t) / (1 + \rho_+)$ . Consistency of the family consumption plan requires that these two ratios be equal and one "reasonable" way to achieve this is to "reverse discount" and to set  $\rho_- = \rho_+ = \rho$ .

must exceed  $n$ :

$$(4) \quad v_t = \sum_{i=-1}^{\infty} \left( \frac{1+n}{1+\rho} \right)^i u(c_{t+i}^1, c_{t+i}^2).$$

Barro and Feldstein, and the other authors cited above, have focused on steady states. Here there will be either gifts from children to parents, or bequests from parents to children, or neither, but not both. It is possible, however, that if government debt or Social Security were introduced into a steady state with gifts, the generation that was old when the plan was introduced would choose to leave bequests.<sup>5</sup> Accordingly, I have set up the individual's budget constraint with both gifts and bequests.<sup>6</sup> I assume individuals work full time when young,<sup>7</sup> and not at all when old. For someone born at time  $t$ , the budget constraint is given by

$$(5) \quad c_t^1 + g_t + \frac{c_t^2}{1+r_{t+1}} + \frac{f_t}{1+r_{t+1}} \\ = w_t + \frac{g_{t+1}(1+n)}{1+r_{t+1}} + \frac{f_{t+1}}{1+n},$$

where  $w_t$  is the wage rate at time  $t$ ,  $r_{t+1}$  is the interest rate at time  $t+1$ ,  $g_t$  is the gift from child to parent at the end of period  $t$ , and  $f_t$  is the bequest left by a person of generation  $t$  to his children at the end of period  $t+1$ . Solving (5) for  $c_t^2$ , substituting into (4), and maximizing with respect to  $c_t^1$ ,  $g_t$ , and  $f_t$ , one obtains

$$(6) \quad (\partial u / \partial c_t^1) / (\partial u / \partial c_t^2) = 1 + r_{t+1},$$

$$(7) \quad \partial u(c_t^1, c_t^2) / \partial c_t^1 \\ = \left( \frac{1+r_{t+1}}{1+\rho} \right) \partial u(c_{t+1}^1, c_{t+1}^2) / \partial c_{t+1}^1, \\ i = t-1, t.$$

<sup>5</sup>An example of this is given in the next section.

<sup>6</sup>Carmichael deals with gifts and bequests separately for reasons of tractability (1982, p. 203, fn. 4).

<sup>7</sup>I discuss the possibility of a labor-leisure choice in Section III.

Equation (6) is the familiar condition that, in equilibrium, the slope of an individual's indifference curve should equal the slope of his budget constraint. Equation (7) states that, in equilibrium, the intergenerational  $MRS(c_t^1, c_{t+1}^1)$ , which discounts  $t+1$  utility at  $\rho$ , must equal the intertemporal price ratio,  $1+r_{t+1}$ . Equations (8)–(13) complete the model.

$$(8) \quad y_t = f(k_t), \quad f' > 0; f'' < 0,$$

$$(9) \quad w_t = f(k_t) - f'(k_t)k_t,$$

$$(10) \quad r_t = f'(k_t),$$

$$(11) \quad h_t = (w_{t-1} + f_{t-2}/(1+n) \\ - c_{t-1}^1 - g_{t-1})/(1+n),$$

$$(12) \quad h_t = k_t,$$

$$(13) \quad c_{t-1}^2 + f_{t-1} = (1+n)(1+r_t)h_t \\ + (1+n)g_t.$$

Equation (8) is the constant-returns-to-scale production function in intensive form, and (9) and (10) are the equilibrium wage and interest rate equations. Define  $h_t$  to be private wealth, per young person, at the beginning of period  $t$ . Thus  $(1+n)h_t$  must equal the assets of those who have just become old—wages, plus bequests, minus consumption when young, minus gifts to their parents. Equation (12) is an identity that states that, in the absence of government (or institutional) assets or debts, private wealth must equal the stock of capital. Equation (13) is an identity that says that for each old person, consumption plus bequests equals wealth, plus interest, plus gifts received from children. Only one of  $g_t$  or  $f_{t-1}$  can be positive.

As observed in the introduction, the Barro model is very similar to optimal growth models. Its steady state revolves around the modified Golden Rule,  $r = \rho > n$ , which follows

from equation (7).<sup>8</sup> This result contrasts with Carmichael's conclusion that "...there are no a priori grounds on which to expect any particular relationship between the growth and interest rates" (1982, p. 202). If one were to ignore equation (4) and assume instead that each generation's utility is described by  $u(c_t^1, c_t^2)$ , there would be a steady-state equilibrium interest rate,  $r^*$ , which could have any relationship to  $n$  (as in Diamond, 1965). If (4) were then reintroduced, with  $\rho$  necessarily greater than  $n$ , the system would asymptotically approach a steady state where  $r$  equals  $\rho$ . If  $\rho$  were greater than  $r^*$ , so that the capital stock has to be reduced in the new steady state, children would make gifts to their parents; if  $r^*$  were greater than  $\rho$ , bequests would occur; if  $r^*$  equalled  $\rho$ , neither would occur.

## II. Government Debt

Suppose that at time 0 the old are each given a lump sum transfer of  $(1+n)^2e$ , and these are financed by issuing  $(1+n)e$  of government debt to each young person at time 0, where  $e$  is the stock of government (internal) debt per young person at time 1. Following Diamond (1965), assume that the government fixes  $e$  by levying a tax on the wage rate of the young at time  $t$ ,  $\theta_t$ . Then the following equation holds.

$$(14) \quad \theta_t w_t = (r_t - n)e, \quad t \geq 1 \quad (\theta_t = 0, t < 1).$$

The only other modifications required in equations (6)–(13) are to multiply  $w_{t-1}$  by  $(1-\theta_{t-1})$  in (11), to subtract  $e$  from the left-hand side of equation (12), and to add  $(1+n)^2e$  to the right-hand side of (13) only for  $c_{-1}^2$ .

It is clear that the vectors  $\{f_t\}_{t=-1}^\infty$  and  $\{g_t\}_{t=0}^\infty$  are sufficient instruments for private

<sup>8</sup>With technical progress in the steady state at rate  $m$ , one obtains

$$1+r = (1+\rho) \frac{\partial u(c_{t-1}^1, c_{t-1}^2)/\partial c_{t-1}^1}{\partial u(\{(1+m)c_{t-1}^1, (1+m)c_{t-1}^2\}/\partial c_t^1}.$$

If  $u$  were homogeneous of degree 1, marginal utilities would be homogeneous of degree 0 and  $r$  would equal  $\rho$ .

individuals to prevent the program from having any real effects. Since each generation was already carrying out the optimal consumption plan, no consumption levels are altered and thus  $k$ ,  $y$ ,  $w$  and  $r$  are unchanged. Let  $z_{t-1}$  equal  $(f_{t-1}/(1+n)) - g_t$ . Equations (11) and (12) imply that  $\Delta z_{t-1} = \Delta((1+n)e_{t+1} + \theta_t w_t)$  for all  $t$ . If  $g$  were positive, initially, and thus  $f$  was zero, the program would induce the young at time 0 to cut  $g$  by  $(1+n)e$ . If  $(1+n)e$  were greater than  $g$ , the new level of  $g$  would be set equal to zero, and each old person at time 0 would increase  $f$  by  $(1+n)((1+n)e - g)$ . If, on the other hand,  $f$  were positive initially,  $\Delta f_{-1}$  would equal  $(1+n)^2e$ . In either case, the change in  $z$  in the next period will be  $(r-n)e$ , and zero thereafter. Similarly, individuals could prevent the introduction of Social Security, or some other government programs, from having any real effects. However, if an interest income tax,  $\tau$ , were introduced into this model, equation (7) would dictate that  $r(1-\tau)$  equalled  $\rho$  in steady-state equilibrium. An increase in  $\tau$ , for any purpose, would imply an increase in  $r$  and hence a reduction in the capital-labor ratio, and each individual's utility level in steady-state equilibrium (see Lawrence Summers, 1981, p. 537).

The model in Diamond (1965) is identical to this one except that individuals do not care about their parents or heirs in his model. Since the effects of Social Security are much like those of government debt, the heart of the Barro-Feldstein debate (1976) is a difference of opinion about individual utility functions. Most readers would experience considerable difficulty in discerning this from the published record.

The results of this section point to a conundrum in Carmichael's article. He states that "...for a restricted Nash concept of equilibrium, the results of the one-sided analysis extend to the more complex two-sided case...[with gifts and bequests]" (p. 203, fn. 4). Suppose government debt is introduced into his version of the Barro model with gifts, where according to his analysis  $n > r$  (p. 206). If the debt is sufficiently large, gifts would have to be set equal to zero and bequests would become positive; but with

bequests, he states that  $r > n$  (p. 205). This means that the economy moves from a steady state where  $r < n$  to one where  $r > n$ ; this implies that his "extended neutrality theorem," expressed in Propositions 1 and 2 (p. 207), does not hold in this case.

The above discussion prompts one further comment. A government might introduce debt into a Diamond model if it valued the utility of current generations more highly than the utility of future generations. By way of contrast, there is no rationale for introducing debt into a Barro model. It will not raise the utility of any generation and may well reduce it, if those who are paying the wage tax are permitted a labor-leisure choice.<sup>9</sup>

### III. A Labor-Leisure Choice in the Barro Model

Let  $l$  represent each individual's leisure when young so that the utility function in equation (4) now includes  $l$  as an argument. The first-order conditions for utility maximization, (6) and (7), are not basically changed, and there will be one additional equation that says that the  $MRS(l, c)$  equals the (after-tax) wage rate. I shall concentrate on the effects of government debt in steady-state equilibrium.

As above,  $r$  must equal  $\rho$  in steady-state equilibrium, and hence, the introduction of debt cannot affect the steady-state levels of  $k$ ,  $y$ , or  $w$ . The following equations are the steady-state, with government debt, versions of (6), (11), (12), (13), (14) and the new  $MRS(l, c)$ , condition.

$$(15) \quad u_{c^1}/u_{c^2} = 1 + \rho,$$

$$(16) \quad u_l/u_{c^1} = (1 - \theta)w,$$

$$(17) \quad h = \{(1 - \theta)(1 - l)w + z - c^1\}/(1 + n),$$

$$(18) \quad h - e = k(1 - l),$$

$$(19) \quad c^2 + (1 + n)z = (1 + n)(1 + \rho)h,$$

$$(20) \quad \theta(1 - l)w = (\rho - n)e.$$

<sup>9</sup>That debt is nonneutral once a labor-leisure choice is permitted is acknowledged, but not analyzed, in Carmichael (1982, p. 211).

It can be shown that  $dl/de$  will be positive, provided the second-order conditions for the problem are satisfied. This means that each young person will work less and individual utility will be lower, with government debt. Even though the capital-labor ratio is unaffected, *the capital stock will be lower*. Clearly debt is not neutral once one permits a labor-leisure choice to the young who pay taxes on their earnings.

### IV. Conclusions

The main purpose of this paper is not to defend the Barro model as an accurate description of the real world, but rather to examine the logical implications of assumptions that he and others have made. I have argued that the Barro model is essentially the optimal growth model, in which the economy exhibits the modified Golden Rule in steady-state equilibrium. If this interpretation is correct, a number of results follow. First, Barro's hunch, which was that inefficient ( $r < n$ ) steady states could be ruled out for his model, is correct. Second, much of the interchange between Barro and Feldstein, which revolves around  $n > r$  steady states, obscures the real issue, which concerns the nature of individual preferences. Third, I believe there is a conundrum in Carmichael's article that results from his (and Buiter's) faulty specification of the individual's optimization problem in the Barro model. Fourth, when a labor-leisure choice is permitted the young, who are taxed to hold constant the stock of debt per young person, government debt is no longer neutral; although it does not affect the capital-labor ratio, it does reduce the capital stock.

### REFERENCES

- Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November-December 1974, 82, 1095-117.  
 ———, "Reply to Feldstein and Buchanan," *Journal of Political Economy*, April 1976, 84, 343-49.  
 Bewley, Truman, (1981a) "The Indeterminacy of Interest Rates," Discussion Paper

- No. 491, Northwestern University, August 1981.
- \_\_\_\_\_, (1981b) "The Relation Between Social Security, Saving and Investment in a Life-Cycle Model," Discussion Paper No. 492, Northwestern University, August 1981.
- Buiter, Willem, "Government Finance in an Overlapping-Generations Model with Gifts and Bequests," in George M. von Furstenberg, ed., *Social Security Versus Private Saving*, Cambridge: Ballinger, 1979.
- Carmichael, Jeffrey, "Economic Equilibrium and Steady-State Growth with Intergenerationally-Dependent Preferences," Research Memo. No. 245, Princeton University, 1979.
- \_\_\_\_\_, "On Barro's Theorem of Debt Neutrality: The Irrelevance of Net Wealth," *American Economic Review*, March 1982, 72, 202-13.
- Cass, David, "Optimum Growth in an Aggregate Model of Capital Accumulation," *Review of Economic Studies*, July 1965, 32, 233-40.
- Diamond, Peter, "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, 55, 1126-50.
- \_\_\_\_\_, "Taxation and Public Production in a Growth Setting," in J. A. Mirrlees and N. H. Stern, eds., *Models of Economic Growth*, IEA Conference at Jerusalem, New York: Wiley, 1973.
- Feldstein, Martin, "Social Security, Induced Retirement and Aggregate Capital Accumulation," *Journal of Political Economy*, September-October, 1974, 82, 905-26.
- \_\_\_\_\_, "Perceived Wealth in Bonds and Social Security: A Comment," *Journal of Political Economy*, April 1976, 84, 331-36.
- \_\_\_\_\_, "Social Security and Private Savings: International Evidence in an Extended Life-Cycle Model," in his and Robert Inman, eds., *The Economics of Public Services*, IEA Conference at Turin, London: Macmillan, 1977.
- Sidrauski, Miguel, "Rational Choice and Patterns of Growth in a Monetary Economy," *American Economic Review Proceedings*, May 1967, 57, 534-44.
- Summers, Lawrence, "Capital Taxation and Accumulation in a Life Cycle Growth Model," *American Economic Review*, September 1981, 71, 533-44.

# Optimal Investment under Uncertainty

By ANDREW B. ABEL\*

This paper examines the effect of output price uncertainty on the investment decision of a risk-neutral competitive firm which faces convex costs of adjustment.<sup>1</sup> This issue has been analyzed by Richard Hartman (1972) and by Robert Pindyck (1982), but they reached dramatically different results. Hartman showed that with a linearly homogeneous production function, increased output price uncertainty leads the competitive firm to increase its investment. However, Pindyck found increased output price uncertainty leads to increased investment only if the marginal adjustment cost function is convex; but, if the marginal adjustment cost function is concave, then increased uncertainty will reduce the rate of investment. Pindyck argues that his results differ from Hartman's results because of a different stochastic specification of the price of output. In Hartman's discrete-time model, price is random in each period including the current period, whereas in Pindyck's continuous-time model, the current price is known but the future evolution of prices is stochastic. In this paper, I demonstrate that Hartman's results continue to hold using Pindyck's stochastic specification and that Pindyck's analysis applies to a so-called "target" rate of investment, which in general is not optimal.

The model developed herein, which is a special case of Pindyck's model, is used because it can be solved explicitly, unlike

Pindyck's more general model. Since Pindyck did not derive an expression for the optimal rate of investment, he used a phase diagram to determine the target capital stock. This target capital stock is determined by the intersection of a locus for which the rate of change of the capital stock is zero, and a locus for which the expected change in the rate of investment is zero. A problem with this stochastic phase diagram approach is that in general there is no reason for the firm to be on the locus with zero expected change in investment, even in the long run. Indeed, in the particular model in this paper, optimal behavior is such that the expected proportional rate of change of investment is (in general, a nonzero) constant over time.

## I. The Model of the Firm

Since the model presented below is a special case of Pindyck's model, the description of it will be brief. The competitive firm uses labor,  $L_t$ , and capital,  $K_t$ , to produce output according to a Cobb-Douglas production function. The firm hires labor at a fixed wage rate  $w$  and undertakes gross investment  $I_t$ , by incurring an increasing convex cost of adjustment  $c(I_t)$ . It is assumed that the cost of adjustment function has constant elasticity  $\beta > 1$ . Therefore, the firm's cash flow at time  $t$  is  $p_t L_t^\alpha K_t^{1-\alpha} - wL_t - \gamma I_t^\beta$  where  $p_t$  is the price of output. Suppose that the firm is risk neutral and maximizes the expected present value of its cash flow subject to the capital accumulation equation

$$(1) \quad dK_t = (I_t - \delta K_t) dt,$$

and the equation which describes the behavior of the price of output

$$(2) \quad dp_t/p_t = \sigma dz,$$

where  $dz$  is a Wiener process with mean zero and unit variance. Equation (1) simply states

\*Harvard University and National Bureau of Economic Research. I thank Ernst Berndt, Stanley Fischer, Robert McDonald, Peter Merrill, Robert Pindyck, and Lawrence Summers for helpful discussions. I also thank the participants in workshops at Columbia University, Harvard University, and MIT for comments on earlier drafts of a longer version of this paper.

<sup>1</sup>Cost of adjustment models were introduced by Robert Eisner and Robert Strotz (1963), Robert Lucas (1967), John Gould (1968) and Arthur Treadway (1969). More recently, Michael Mussa (1977), my (1979, 1981, 1982) studies, Hiroshi Yoshikawa (1980), and Fumio Hayashi (1982) have used cost of adjustment models to provide a more rigorous foundation for James Tobin's (1969)  $q$  theory of investment.

that net investment is equal to gross investment less depreciation where  $\delta$  is the constant proportional rate of physical depreciation. The price process described by (2) has the properties<sup>2</sup> that  $E_t(p_s) = p_t$ ,  $s \geq t$ , and the variance of  $p_s$ , conditional on  $p_t$ , is  $(s - t)\sigma^2$ . The value of the firm is the maximized expected present value of cash flow. Assuming that the discount rate  $r$  is constant, we can write the value of the firm as

$$(3) \quad V(K_t, p_t) = \max_{I_t, L_t} E_t \int_t^\infty [p_s L_s^\alpha K_s^{1-\alpha} - wL_s - \gamma I_s^\beta] \exp(-r(s-t)) ds,$$

where the maximization is subject to the constraints in (1) and (2).

The value function in (3) must obey the following optimality condition

$$(4) \quad rV(K_t, p_t) dt = \max_{I_t, L_t} [p_t L_t^\alpha K_t^{1-\alpha} - wL_t - \gamma I_t^\beta] dt + E_t(dV).$$

The optimality condition in (4) has a straightforward economic interpretation. If the owners of the firm require a mean rate of return  $r$ , then the left-hand side of (4) is the total mean return required by the owners of the firm over the time interval  $dt$ . The right-hand side of (4) is the total return expected by the owners of the firm. It consists of the cash flow plus the expected capital gain or loss  $E_t(dV)$ . Optimality requires that the expected return equals the required mean return.

To calculate the capital gain or loss,  $dV$ , we recognize that the value of the firm is a function of the two state variables  $K_t$  and  $p_t$  and then apply Ito's Lemma to obtain

$$(5) \quad dV = V_K dK + V_p dp + (1/2)V_{KK}(dK)^2 + (1/2)V_{pp}(dp)^2 + V_{pK}(dp)(dK).$$

<sup>2</sup>For good discussions of stochastic calculus set in an economic context, the reader is referred to William Brock, Gregory Chow (1981), Stanley Fischer (1975), and Robert Merton (1971). The solution to a more general form of the stochastic differential equation in (2) is presented in Fischer, equation (13A).

Substituting (1) and (2) into (5), and recognizing that  $E_t(dz) = (dt)^2 = (dt)(dz) = 0$ , we obtain the expected change in the value of the firm over the time interval  $dt$ :

$$(6) \quad E_t(dV) = [(I_t - \delta K_t)V_K + (1/2)p_t^2 \sigma^2 V_{pp}] dt.$$

Substituting (6) into (4) yields

$$(7) \quad rV(K_t, p_t) = \max_{L_t, I_t} \{p_t L_t^\alpha K_t^{1-\alpha} - wL_t - \gamma I_t^\beta + (I_t - \delta K_t)V_K + \frac{1}{2}p_t^2 \sigma^2 V_{pp}\}.$$

It is easily shown that

$$(8) \quad \max_{L_t} \{p_t L_t^\alpha K_t^{1-\alpha} - wL_t\} = hp_t^{1/(1-\alpha)} K_t,$$

where  $h = (1-\alpha)(\alpha/w)^{\alpha/(1-\alpha)}$ .

Observe that  $hp_t^{1/(1-\alpha)}$  is the marginal revenue product of capital.

Differentiating the right-hand side of (7) with respect to  $I_t$ , we obtain

$$(9) \quad \gamma \beta I_t^{\beta-1} = V_K.$$

According to (9), the optimal rate of investment is such that the marginal cost of investment is equal to the marginal valuation of capital  $V_K$ . Substituting (8) and (9) into (7) yields

$$(10) \quad rV(K_t, p_t) = hp_t^{1/(1-\alpha)} K_t + (\beta - 1)\gamma I_t^\beta - \delta K_t V_K + (1/2)p_t^2 \sigma^2 V_{pp}.$$

Equations (9) and (10) together can be expressed as a nonlinear second-order partial differential equation. In general, such equations cannot be solved explicitly, as noted by Pindyck. However, I have imposed enough structure on this problem to obtain an explicit solution. It can be verified that the

equations below satisfy (9) and (10).

(11a)

$$V(K_t, p_t) = q_t K_t + \frac{(\beta - 1)\gamma(q_t/\beta\gamma)^{\beta/(\beta-1)}}{r - \frac{\beta(1-\alpha+\alpha\beta)\sigma^2}{2(1-\alpha)^2(\beta-1)^2}}$$

where

$$(11b) \quad q_t = \frac{hp_t^{1/(1-\alpha)}}{r + \delta - \frac{\alpha\sigma^2}{2(1-\alpha)^2}}$$

and

$$(12) \quad I_t = (q_t/\beta\gamma)^{1/(\beta-1)}.$$

Several results follow immediately from equations (11a), (11b), and (12). First we observe that the value of the firm is a linear function of the capital stock, since the slope of the value function,  $q_t$ , is independent of the capital stock.<sup>3</sup> As shown in Section II,  $q_t$  is equal to the present value of expected marginal revenue products of capital. Since, for a competitive firm with a constant returns to scale production function, the marginal product of capital depends only on the real wage rate, and thus is independent of the level of the capital stock, it follows that  $q_t$  is independent of  $K_t$ . According to (12), the optimal rate of investment is an increasing function of  $q_t$ . Moreover,  $I_t$  depends only on  $q_t$  and is independent of  $K_t$ .

## II. The Effect of Uncertainty of Investment

Since the optimal rate of investment is an increasing function of  $q_t$ , and depends only on  $q_t$ , we can determine the qualitative effect of uncertainty on investment simply by analyzing the effect of uncertainty on  $q_t$ . It follows immediately from (11b) that for a given level of the current price of output  $p_t$ , an increase in uncertainty, as measured by  $\sigma^2$ , will lead to an increase in the optimal rate of investment. Contrary to the results of

Pindyck, this result holds whether the marginal adjustment function is convex ( $\beta > 2$ ), concave ( $\beta < 2$ ) or linear ( $\beta = 2$ ).

To explain the positive effect of uncertainty on investment, I will first show that  $q_t$  is the expected present value of marginal revenue products accruing to the undepreciated portion of capital from time  $t$  onward. Since the marginal revenue product of capital,  $p_t F_{K_t}$ , is equal to  $hp_t^{1/(1-\alpha)}$ , it can be shown that, for the price process in (2),<sup>4</sup>

$$(13) \quad E_t(p_s F_{K_s}) = hE_t(p_s^{1/(1-\alpha)}) \\ = hp_t^{1/(1-\alpha)} \exp[\alpha\sigma^2(s-t)/2(1-\alpha)^2].$$

Using (13), the expected present value of marginal revenue products of capital is

$$(14) \quad \int_t^\infty E_t(p_s F_{K_s}) \exp[-(r+\delta)(s-t)] ds \\ = \int_t^\infty hp_t^{1/(1-\alpha)} \exp\left[\left(\alpha\sigma^2(s-t) \right. \right. \\ \left. \left. /2(1-\alpha)^2 - (r+\delta)(s-t)\right) \right] ds.$$

The integral on the right-hand side of (14) can be evaluated by inspection and is obviously equal to  $q_t$  in (11b). Thus  $q_t$  is indeed the expected present value of marginal products of capital. Note from equation (13) that increased uncertainty tends to increase the expected value of future marginal revenue products of capital and hence increases  $q_t$  and investment. Although equation (13) applies only for a Cobb-Douglas production function, the reasoning applies more generally to competitive firms with linearly homogeneous production functions. As long as the marginal revenue product of capital is a strictly convex function of the price of output, then increased uncertainty about the future price of output tends to increase the

<sup>3</sup>Mussa showed that for a linearly homogeneous production function  $F(K, L)$ , the value of the firm under certainty is linear in  $K_t$ .

<sup>4</sup>Given  $p_t$ , the log of the price of output at some future date  $s$  is normally distributed with  $E_t(\ln p_s) = \ln p_t - (1/2)\sigma^2(s-t)$ , and  $\text{var}_t(\ln p_s) = \sigma^2(s-t)$  (see Fischer's Appendix). Using the fact that if  $\ln x$  is normally distributed, then  $E(x) = \exp[E(\ln x) + (1/2)\text{var}(\ln x)]$ , we can derive my equation (13).

expected future marginal revenue product, and hence increases both  $q_t$  and investment.<sup>5</sup>

Contrary to the results presented above, Pindyck finds that the effect of uncertainty on investment depends on the curvature of the marginal adjustment cost function. His results are derived under the assumption that (eventually) the expected rate of change of investment,  $E_t(dI_t)/dt$ , is equal to zero. However, the optimal rate of investment does not, in general, obey this assumption.

To examine the dynamic behavior of investment, I first apply Ito's Lemma to (11a) to obtain

$$(15) \quad \frac{dq_t}{q_t} = \frac{1}{1-\alpha} \frac{dp_t}{p_t} + \frac{\alpha}{2(1-\alpha)^2} \left( \frac{dp_t}{p_t} \right)^2,$$

which implies

$$(16) \quad (1/dt)E_t(dq_t/q_t) = \frac{\alpha\sigma^2}{2(1-\alpha)^2}.$$

Substituting (16) into (11b), we obtain

$$(17) \quad q_t = hp_t^{1/(1-\alpha)} / (r + \delta - E_t(dq_t/q_t)/dt).$$

Interpreting  $q_t$  as the shadow price of capital, the user cost of capital is  $[r + \delta - (1/dt)E_t(dq_t/q_t)]q_t$ . Therefore, equation (17) merely expresses the equality of the marginal revenue product of capital and the user cost of capital.

Now to analyze the dynamic behavior of investment, let us apply Ito's Lemma to (12) to obtain

$$(18) \quad \frac{dI_t}{I_t} = \frac{1}{\beta-1} \frac{dq_t}{q_t} + \frac{2-\beta}{2(\beta-1)^2} \left( \frac{dq_t}{q_t} \right)^2.$$

Taking expectations on both sides of (18), and using (15) to calculate  $(dq_t/q_t)^2$ , we

obtain

$$(19) \quad \frac{1}{dt}E_t\left(\frac{dI_t}{I_t}\right) = \frac{1}{(\beta-1)} \frac{1}{dt}E_t\left(\frac{dq_t}{q_t}\right) + \frac{(2-\beta)\sigma^2}{2(\beta-1)^2(1-\alpha)^2}.$$

Now substituting (16) into (19) yields

$$(20) \quad \frac{1}{dt}E_t\left(\frac{dI_t}{I_t}\right) = \frac{1}{2(\beta-1)(1-\alpha)^2} \left( \alpha + \frac{2-\beta}{\beta-1} \right) \sigma^2.$$

From equation (20), we observe that the expected proportional growth rate of investment is independent of the state variables and is constant over time. Although this constant growth rate is zero under certainty ( $\sigma^2 = 0$ ), we find that in the presence of uncertainty, the expected growth rate of investment is not equal to zero in general, nor does it tend toward zero. Thus Pindyck's analysis, which assumes that  $E_t(dI_t) = 0$ , is inappropriate to the analysis of the behavior of the optimal rate of investment.<sup>6</sup>

### III. Concluding Comments

Pindyck has emphasized the curvature of the marginal adjustment cost function in determining the effect of uncertainty on investment. Although I have shown that, given the current price of output, higher uncertainty leads to a higher current rate of investment regardless of the curvature of the marginal adjustment cost function, this curvature does

<sup>5</sup>This line of argument was developed by Richard Hartman (1972).

<sup>6</sup>In order for Pindyck's analysis to apply to optimal investment behavior, the expression on the right-hand side of (20) must equal zero. This expression is zero if either (a) there is no uncertainty ( $\sigma = 0$ ) or (b) the parameters of technology happen to be such that  $\alpha = (\beta - 2)/(\beta - 1)$ . More generally, if the price of output evolves according to  $dp_t/p_t = \pi dt + \sigma dz$ , where  $\pi$  is the expected rate of inflation, it can be shown that the expected rate of change of optimal investment is zero if and only if  $\pi = [\alpha + (2 - \beta)/(\beta - 1)]\sigma^2/2(1 - \alpha)$ . (See my 1981 paper.) Pindyck's results apply only to situations in which this condition holds.

have an important implication for the relation between the expected growth rate of investment and the expected growth rate of the marginal valuation of capital,  $q_t$ . Under certainty, the growth rate of investment is equal to the growth rate of  $q_t$  multiplied by the elasticity of investment with respect to  $q_t$ ,  $1/(\beta - 1)$ , as may be verified from (19). However, under uncertainty, this relation holds only if the marginal adjustment cost function is linear. If the marginal adjustment cost is convex (concave), then, under uncertainty, the expected growth rate of investment is less (greater) than the expected growth rate of  $q_t$  multiplied by the elasticity of investment with respect to  $q_t$ .

The analysis of this paper is easily extended to allow for uncertainty in the wage rate,  $w$ , and uncertainty in  $\gamma$ , which enters multiplicatively into the adjustment cost function. In this extended framework, the value function is again linear in the capital stock. Investment is an increasing function of only  $q_t/\gamma_t$ , where  $q_t$  is the slope of the value function.<sup>7</sup> Uncertainty affects investment only to the extent that it affects the variance of the logarithm of the real wage rate. Specifically, increased variance in the real wage rate leads to an increase in the optimal rate of investment.

Finally, note that, according to (16), the marginal valuation of capital  $q_t$  is expected to grow without bound as we look further and further into the future. This disquieting feature of the model is a consequence of the assumption in (2) that  $p_t$  evolves according to a random walk. Therefore, given today's price  $p_t$ , the variance of the future price of output,  $p_s$ , grows without bound as  $s$  grows without bound. Since the marginal revenue product of capital is a convex function of the price of

output, the expected value of this marginal revenue product is an increasing function of the variance of the price. Therefore, the expected marginal revenue product grows without bound over time. This feature of the model could be removed by assuming that the price of output evolves according to a process for which the forecast variance is bounded. However, in the present context, the easy interpretations of the explicit solutions made possible by the random walk assumption seem to be worth the cost.

## REFERENCES

- Abel, Andrew B., *Investment and the Value of Capital*, New York: Garland Publishing Co., 1979.
- , "Optimal Investment Under Uncertainty: Towards a Stochastic  $q$  Theory," Discussion Paper No. 873, Harvard Institute of Economic Research, July 1981, rev. December 1981.
- , "Dynamic Effects of Permanent and Temporary Tax Policies in a  $q$  Model of Investment," *Journal of Monetary Economics*, May 1982, 9, 353–73.
- Brock, William A., "Introduction to Stochastic Calculus: A User's Manual," mimeo., University of Chicago.
- Chow, Gregory C., *Econometric Analysis by Control Methods*, New York: John Wiley and Sons, 1981.
- Eisner, Robert and Strotz, Robert, "Determinants of Business Investment," in *Impacts of Monetary Policy*, Englewood Cliffs: Prentice-Hall, 1963, 59–337.
- Fischer, Stanley, "The Demand for Index Bonds," *Journal of Political Economy*, June 1975, 83, 509–34.
- Gould, John P., "Adjustment Costs in the Theory of Investment of the Firm," *Review of Economic Studies*, January 1968, 35, 47–55.
- Hartman, Richard, "The Effects of Price and Cost Uncertainty on Investment," *Journal of Economic Theory*, October 1972, 5, 258–66.
- Hayashi, Fumio, "Tobin's Marginal and Average  $q$ : A Neoclassical Interpretation," *Econometrica*, January 1982, 50, 213–24.
- Lucas, Robert E., Jr., "Adjustment Costs and

<sup>7</sup>If  $dp_t/p_t = \pi_p dt + \sigma_p dz_p$ ,  $dw_t/w_t = \pi_w dt + \sigma_w dz_w$  and  $d\gamma_t/\gamma_t = \pi_\gamma dt + \sigma_\gamma dz_\gamma$  where  $dz_p$ ,  $dz_w$ , and  $dz_\gamma$  are Wiener processes with mean zero and unit variance, then the optimal rate of investment is proportional to  $(q_t/\gamma_t)^{1/(\beta-1)}$  where

$$q_t = \frac{hp_t^{1/(1-\alpha)}}{r + \delta - \frac{1}{1-\alpha}[\pi_p - \alpha\pi_w] - \frac{1}{2} \frac{\alpha}{(1-\alpha)^2} \text{var}(p-w)}$$

See my 1981 paper for details.

- the Theory of Supply," *Journal of Political Economy*, August 1967, 75, 321-34.
- Merton, Robert C., "Optimum Consumption and Portfolio Rules in a Continuous-Time Model," *Journal of Economic Theory*, December 1971, 3, 373-413.
- Mussa, Michael, "Market Value and the Investment Decision in an Adjustment Cost Model of Firm Behavior," Discussion Paper No. 74-15, Department of Economics, University of Rochester, July 1974.
- \_\_\_\_\_, "External and Internal Adjustment Costs and the Theory of Aggregate and Firm Investment," *Economica*, May 1977, 47, 163-78.
- Pindyck, Robert S., "Adjustment Costs, Uncertainty, and the Behavior of the Firm," *American Economic Review*, June 1982, 72, 415-27.
- Tobin, James, "A General Equilibrium Approach to Monetary Theory," *Journal of Money, Credit and Banking*, February 1969, 1, 15-29.
- Treadway, Arthur B., "On Rational Entrepreneurial Behavior and the Demand for Investment," *Review of Economic Studies*, April 1969, 36, 227-39.
- Yoshikawa, Hiroshi, "On the 'q' Theory of Investment," *American Economic Review*, September 1980, 70, 739-43.

# Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?: Comment

By BASIL L. COPELAND, JR.\*

Robert Shiller's analysis of the volatility of stock prices in a recent article in this *Review* led him to conclude that stock prices are too volatile to be explained by subsequent changes in dividends. A second possible explanation, that the volatility of stock prices reflects real changes in the investor discount rate, is also thought to be inadequate to account for the volatility of stock prices. Shiller's findings would thus seem to call into question the validity of the efficient markets model of stock valuation. The purpose of this comment is to point to a third hypothesis that would not only explain the volatile nature of stock prices, but that is, in fact, a predictable consequence of the efficient markets model.

Shiller's analysis focuses upon transitory shocks that are due to unanticipated dividend changes. However, dividend changes may also represent fundamental changes in a firm's retention policies which would affect the dividend growth rate. If investors react to dividend changes as if they are changes in what investment analysts sometimes call "fundamentals," then stock prices will be far more volatile than if the change is viewed as transitory. The difference between a transitory change and a fundamental change is, if you will, fundamental. Figure 1 illustrates the difference between a transitory change and a fundamental change as I use the terms. A transitory change represents a one-time shock in the path along which dividends grow; the rate of growth itself is not affected. A fundamental change represents a permanent change in the slope of the path along which dividends grow. Price volatility induced by changes in dividends depends upon whether investors view the dividend change as transitory or permanent.

To compare the impact of the two types of change on stock price volatility, consider the

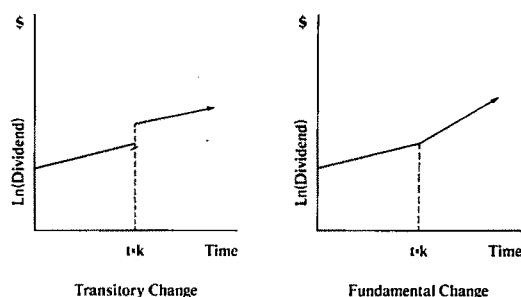


FIGURE 1

following example of a one-period change in the dividend. Following Shiller, the one-period holding return  $H_t \equiv (P_{t+1} + D_t)/P_t$  is the return from buying the stock at time  $t$  and selling it at time  $t+1$ , where  $P_t$  is the price of the stock at time  $t$ ,  $\Delta P_{t+1}$  is the one-period change in price, and  $D_t$  is the dividend receipt in period  $t$ . The investor discount rate,  $r$ , equals the equilibrium expected holding period return:  $E(H_t) = r$ . To begin, suppose that  $r = 0.06$ , that the dividend at time  $t$  is \$1.00, and that it is expected to increase to \$1.03 at time  $t+1$ . The expected growth rate is 0.03, and the stock will be priced to yield 0.03 on the basis of the current dividend. We can easily calculate that  $P_t = \$33.33$  and that  $\Delta P_{t+1} = \$1.00$ . On an initial investment of \$33.33, the expected return will be 0.06, or \$2.00—a \$1.00 dividend plus a \$1.00 capital gain.

Now suppose that at time  $t+1$ , the dividend that is paid out is \$1.04 instead of the \$1.03 that was expected. What impact will this have on the stock price? The answer to this question depends upon whether investors view the change as transitory or fundamental. Let us consider first the case of a transitory change. When the dividend of \$1.04 is paid out at time  $t+1$ , the equilibrium price will increase to \$34.67 so as to provide a current yield of 0.03, an increase of \$0.34 over the price of \$34.33 that would

\*Senior Consultant, Hess & Lim, Inc.

have resulted with a dividend of \$1.03. In round numbers, a 1 percent change in the dividend (from \$1.03 to \$1.04) would be accompanied by a 1 percent change in the equilibrium price. However, investors may interpret the new dividend as signaling a fundamental change in the rate of growth in the dividend from 3 to 4 percent per year. If that is the case, the dividend yield will fall to 0.02 and the price of the stock will rise to \$52.00 so as to yield 0.02 on the basis of the \$1.04 dividend. In this case, a 1 percent increase in the dividend results in a 50 percent increase in price compared to the \$34.67 that would have resulted in the absence of a fundamental change.

As this simple example shows, the volatility of stock prices largely depends upon how investors react to dividend changes. If the dividend change is considered transitory, as I use the term, then the associated price change will be minimal. This is the type of change that Shiller's analysis models; it models the impact of subsequent changes in the dividend, not the impact of subsequent changes in the *rate of growth* in the dividend. If a

change in the dividend is viewed by investors to signal a fundamental change in the rate of growth in the dividend, even a small change in the dividend can lead to large swings in price. This is because a small change in the dividend can imply large changes in the rate of growth in the dividend. In the example above, a dividend increase of 1 percent, from \$1.03 to \$1.04, implied a 33 percent increase in the dividend growth rate, from 0.03 to 0.04. More importantly, this 33 percent increase in the dividend growth rate leads to a 33 percent reduction in the current yield required, from 0.03 to 0.02, and a 50 percent rise in price. Large swings in price are thus a predictable consequence of the efficient markets model if investors frequently change their expectations regarding the future rate of growth in dividends.

#### REFERENCES

- Shiller, Robert J., "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, 71, 421-36.

## Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?: Reply

By ROBERT J. SHILLER\*

Let me first rephrase Basil Copeland's point in simple growth model terms. If dividends are expected to grow forever at rate  $g$  and the discount rate is  $r > g$ , then  $P_t = D_t / (r - g)$ . A "transitory" change in the growth rate between  $t$  and  $t + 1$  causes dividends to change from the  $D_{t+1}$  that was expected at  $t_0$  to  $D'_{t+1}$ , leaving expected future growth rates unchanged. This will cause  $P'_{t+1}$  to be  $D'_{t+1} / (r - g)$ , so a 1 percent change from the expected  $D_{t+1}$  to  $D'_{t+1}$  causes only a 1 percent change in price. If, on the other hand, the growth rate ( $g'$ ) between  $t$  and  $t + 1$  is "fundamental" and is expected to continue forever, then  $P'_{t+1} = D'_t / (r - g')$ . The change in price could be much larger than the change in dividends. Indeed, by this assumption, the price would change to infinity in any period in which the growth of dividends exceeded the discount rate!

Copeland describes his alternative in terms of investor expectations rather than in terms of dividend processes. Since I presume he is not proposing that the expectations are irrational, I assume he is proposing that the dividend process is such as to make such expectations optimal. By this interpretation, Copeland is giving one expression to the most common criticism I have encountered in connection with my paper: the possibility that uncertainty about future dividends is much higher than suggested by the sample variance of dividends around their historical trend. Copeland is suggesting that the dividend process is unstationary so that the population variance is infinite and the estimated trend is spurious.

Since discussion has centered so much on the variance measurement issue, I infer that these critics have conceded what I thought was a very important point of the paper: that stock price movements cannot be attributed

to information about movements in dividends around the apparent historical trend. Thus, for example, a particular decline in the stock market cannot be interpreted as the result of new information that a recession is on the horizon which will cause a temporary decline in dividends. Even the Great Depression of the 1930's was far too small and transient in its effect on dividends to justify anything like the movements in stock prices regularly observed. Many people seemed not to know this before I wrote this paper, while on the other hand, it is perhaps a rather obvious point and it must have been known to many who pondered the data. A contribution of the paper was to show that given the sort of stationarity assumption commonly implicitly made, conventional regression tests of market efficiency (which consist of regressions of excess returns on information) may fail to reveal a glaring failure of market efficiency. That is, the tests may not be powerful against certain alternatives involving excess volatility. I pointed out in my 1981b paper that these conventional tests do not dominate other tests as some seem to suppose, and that volatility tests may be more robust to data errors.

Perhaps the excess volatility argument might be presented more convincingly in terms of some nonlinear measure of uncertainty of dividends other than their standard deviation around historical trend, say in terms of some measure of dispersion of the percentage change in dividends, but such alternatives seem to be very messy technically. As is usual, we are constrained by our methodology to simple linear models and must learn what we can from them. Of course, we do not literally believe with certainty all the assumptions in the model which are the basis of testing. I did not intend to assert in the paper that I knew dividends were indeed stationary around the historical trend.

\*Cowles Foundation, Yale University.

If, however, one looks at the plot of the real dividend series  $D_t$  (in levels, not detrended) since 1871, one does get the impression that it is dominated by a growth trend and that deviations from the trend tend to reverse themselves. (The real Standard and Poor dividend series which I used in my 1981a paper is plotted in my 1981b paper.) If  $\log(D_t)$  is regressed on  $\log(D_{t-1})$ , a constant and a linear time trend for 1872 to 1978, the coefficient of  $\log(D_{t-1})$  is 0.807 with an estimated standard error of 0.058. If this regression were used to forecast dividends, then  $\log$  dividends would always be expected to return half way to the trend line in three years. This suggests much more stationarity, in a sense, than either the transitory or fundamental stories of Copeland. The transitory story would seem to be represented by the assumption that  $\log$  dividends are a random walk, so that the coefficient of  $\log D_{t-1}$  should be 1.00. Using a small sample test based on Monte Carlo experiments done originally by David Dickey (as reported by Wayne Fuller), we can reject a random walk at the 5 percent level in favor of stationary fluctuations around a trend (see my 1981b paper).

Even if we assumed  $\log$  dividends were a random walk with trend with independent increments, stock prices still would show too much volatility. Assuming  $E_t(D_{t+1}/D_t) = 1 + g$  for all  $t$  and that  $D_{t+1}/D_t$  is independent of  $D_{t'+1}/D_{t'}$ ,  $t \neq t'$ , then the model implies that the price of a share is given by  $P_t = D_t/(r - g)$ . As long as  $g < r$ , price is finite and then the standard deviation of  $P_{t+1}/P_t$  equals the standard deviation of  $D_{t+1}/D_t$ . In fact, with the real Standard and Poor data used in the paper for 1872–1979,  $\hat{\sigma}(P_{t+1}/P_t) = 0.176$  which is greater than  $\hat{\sigma}(D_{t+1}/D_t) = 0.127$ . The ratio of sample variances is 1.93, which by a conventional  $F$  test is significant at the 1 percent level. However, this particular model can be rejected with certainty since the dividend price ratio is not constant.

One can never prove that the dividend process (or some transformation of it) is

stationary. Other forms of nonstationarity (for which Dickey's test is useless) would also invalidate the variance bounds tests. For example, it is conceivable that investors were concerned about possible nationalization of stocks which might cause dividends to fall abruptly to zero. The fact that nationalization didn't chance to happen in the sample doesn't prove that it couldn't happen.

Most of us will probably agree that fear of nationalization is not a likely source for most of the variability of U.S. stock prices since 1871. The challenge for advocates of the efficient markets model is to tell a convincing story which is consistent both with the observed trendiness of dividends for a century and with the high volatility of stock prices. They can certainly tell a story which is within the realm of possibility, but it is hard to see how they could come up with inspiring evidence for the model. The near-total lack of correspondence, except for trend, between the aggregate stock price and its *ex post* rational counterpart (as shown in Figure 1 of my 1981a paper) means that essentially no observed movements in aggregate dividends were ever correctly forecast by movements in aggregate stock prices!

## REFERENCES

- Copeland, Basil L., Jr., "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?: Comment," *American Economic Review*, March 1983, 73, 234–35.
- Fuller, Wayne A., *Introduction to Statistical Time Series*, New York: John Wiley & Sons, 1976.
- Shiller, Robert J., (1981a) "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, 71, 421–36.
- , (1981b) "The Use of Volatility Measures in Assessing Market Efficiency," *Journal of Finance*, May 1981, 36, 291–304.

# Duopoly Models with Consistent Conjectures: Comment

By COLDWELL DANIEL III\*

Timothy Bresnahan's version of the consistent conjectures condition for an oligopolistic equilibrium, recently presented in this *Review*, differs from my own (1970, ch. 8) insofar as his, which will be referred to henceforth as the *BCCE*, requires an equality among the conjectural variations and the slopes of the reaction functions, whereas my own, which will be referred to henceforth as the *DCCE*, requires only that the conjectural variations be such that the reaction functions are consistent. In particular, in the *BCCE* (equations (12), (13), and (14) in Bresnahan), a consistent conjectures equilibrium for a duopoly is defined as a pair of quantities  $q^*$ , and a pair of conjectural variations  $[r_{12}(q_1), r_{21}(q_2)]$ , such that

$$(1) \quad q_1^* = \rho_1(q_2^*), q_2^* = \rho_2(q_1^*),$$

and there is some  $\varepsilon > 0$ , such that

$$(2) \quad r_{12}(q_1) = \partial \rho_2(q_1) / \partial q_1$$

for all  $q_1^* - \varepsilon < q_1 < q_1^* + \varepsilon$ ;

$$(3) \quad r_{21}(q_2) = \partial \rho_1(q_2) / \partial q_2$$

for all  $q_2^* - \varepsilon < q_2 < q_2^* + \varepsilon$ ,

where  $r_{12}(q_1)$  and  $r_{21}(q_2)$  are the conjectural variations of duopolists one and two, respectively, and  $\rho_2(q_1)$  and  $\rho_1(q_2)$  are the reaction functions of duopolists two and one, respectively. In the *DCCE*, a consistent conjectures equilibrium exists if and only if  $r_{12}(q_1)$  and  $r_{21}(q_2)$  are such that  $\rho_1(q_2)$  and  $\rho_2(q_1)$  are consistent in the ordinary simultaneous-equation sense. It will be argued here that, whatever may be the merits of the *BCCE* as a criterion for the local uniqueness of a conjectural variations equilibrium, it is deficient vis-à-vis the *DCCE* as a condition for such an equilibrium, since the *BCCE* is not a

necessary condition for an equilibrium within the parameters of the conventional conjectural variations approach.

In the crucial case of the Cournot model, its solution (equilibrium) is rejected by the *BCCE* because of the incorrectness of the model's conjectural variations (a term which, incidentally, is attributable to Ragnar Frisch, 1933). Two issues are involved here: 1) Must particular agents be required to have prior perfect knowledge under uncertain market conditions in order for there to be a market equilibrium? 2) If not, must particular agents then be required to acquire perfect knowledge (or the equivalent thereof)? The answer to the first question is obviously in the negative, since prior perfect knowledge is the contradictory of uncertainty. The ingeniousness of the Cournot model is that it identifies, in an analytically tractable way, the source of an uncertainty in a situation where two agents, and those two agents exclusively, are in "competition" with each other in the sense that they must act "independently," even though both may know that greater profits are obtainable for each through an agreement. The lack of an equilibrium requirement of a minimal level of profits above zero for either or both duopolists is intentional, relates the first issue cited above to the second, and is important in order for the conjectural variations approach to claim validity as a description of actual free-market processes.<sup>1</sup>

The second question posed above must also be answered in the negative. In general, microeconomic theory does not require particular agents to acquire perfect knowledge,

<sup>1</sup> Its importance in this regard is underscored by Cournot in the following statement: "We say each *independently*, and this restriction is very essential, as will soon appear; for if they should come to an agreement so as to obtain for each the greatest possible income, the results would be entirely different, and would not differ, so far as consumers are concerned, from those obtained in treating of a monopoly" (1960, pp. 79-80).

\*Professor of economics, Memphis State University.

that is, to be correct, in order for a market to be in equilibrium. Consider in this regard the equilibrium of a monopolist who practices limit pricing. The equilibrium is in no way conditioned upon the monopolist acquiring perfect knowledge with respect to the threat of entry. An equilibrium exists whether or not the monopoly is actually threatened with entry. Moreover, an equilibrium exists even though the monopolist's conjecture and his behavior based thereupon is correct and, nevertheless, the monopoly is eventually displaced.

Consider next an equilibrium in a perfectly competitive market for a consumer's good. Since contracts are made for exchange prior to the occurrence of consumption, the equilibrium is based upon allocations of the incomes of consumers in accordance with the conjectured marginal utility per unit of expenditure on each good. The equilibrium exists whether or not the conjectures of the consumers are correct, that is, whether or not the conjectured marginal utilities are those which are realized.

The acquisition of perfect knowledge by a firm is required for an equilibrium only when market conditions are such that the behavior of the firm in equilibrium must be based on perfect knowledge if the firm is to survive. It was Cournot's intention that his duopolists need not acquire perfect knowledge in order to survive in equilibrium.<sup>2</sup> That is, neither prior perfect knowledge nor the acquisition of perfect knowledge is required for the survival of Cournot's duopolists (oligopolists)—being correct, that is, the *BCCE*, is not a necessary condition for an equilibrium within the parameters of the conjectural variations approach.

Bresnahan (see his Sections IV and VI) properly relates conjectural variations equilibria to the matter of the stability of market

structures, as is done traditionally. By removing a free-market barrier to entry, viz, the control of an essential raw material, Cournot describes a transition from monopoly to "unlimited competition" (1960, chs. 5, 7, and 8). Using a conjectural variations approach to analyze oligopolistic markets, H. von Stackelberg concludes that "...we are justified in considering duopoly [oligopoly] as an *unstable market form*" (1952, p. 202). But Stackelberg is right for the wrong reason, since he predicts that "...duopoly tends to lead to the formation of a monopoly" (p. 203). If oligopolistic profits exist and if markets are free, that is, if there are no barriers attributable to governmental intervention, entry can be expected to occur (see my 1970 study, pp. 175-76, and 189; and my 1980 paper). It is entry that brings about competitive or average cost pricing and not the conjectures in conjunction with the form of the marginal cost functions of oligopolists, as is claimed in Bresnahan's (p. 943) intuitive theory of competition (see my 1982 paper).

## REFERENCES

- Bresnahan, Timothy F., "Duopoly Models with Consistent Conjectures," *American Economic Review*, December 1981, 71, 934-45.
- Cournot, A., *Researches into the Mathematical Principles of the Theory of Wealth*, New York: A. M. Kelley, 1960.
- Daniel, Coldwell III, *Mathematical Models in Microeconomics*, Boston: Allyn and Bacon, 1970.
- \_\_\_\_\_, "Windfall Profits: Their Causes and Effects," *Mid-South Quarterly Business Review*, July 1980, 18, 7-11.
- \_\_\_\_\_, "The Effects of Dynamic Free-Market Processes on Product-Market Structures," *Nebraska Journal of Economics and Business*, Autumn 1982, 21, 3-18.
- Frisch, Ragnar, "Monopole-Polypole" *National Økonomisk Tidsskrift*, 1933; cited in R. G. D. Allen, *Mathematical Analysis for Economists*, London: Macmillan, 1953, p. 203.
- von Stackelberg, H., *The Theory of the Market Economy*, London: William Hodge, 1952.

<sup>2</sup>According to Cournot, "...in the moral sphere men cannot be supposed to be free from error and lack of forethought any more than in the physical world bodies can be considered perfectly rigid, or supports perfectly solid, etc." (1960, p. 83).

# Duopoly Models with Consistent Conjectures: Reply

By TIMOTHY F. BRESNAHAN\*

Coldwell Daniel's comment raises both substantive issues and questions of attribution or labelling. The substantive issues miss the point: there are many mathematically correct formulations of oligopoly equilibrium, one of which must be selected by an economic criterion. The questions of labelling are merely misleading.

## I. Labels

Daniel relabels the consistent conjectures equilibrium (*CCE*) a Bresnahan *CCE* (*BCCE*) and labels any oligopoly equilibrium, including those with inconsistent conjectures, a Daniel *CCE* (*DCCE*). In the *DCCE*, that conjectures are "consistent" means that equilibrium exists—the word consistent is completely superfluous. In my paper, "consistency" was a way to select a particular oligopoly solution concept. In William Fellner's language, all oligopolists are right, but only *CCE* oligopolists are right for the right reason.

Since several papers independently took up closely related consistency notions (see citations in my 1981b article), I do not think my name should be added to the *CCE*. As to the *DCCE*, I date the notion that oligopolists' conjectures might be arbitrary well before 1970. Any reader of A. L. Bowley (1924, p. 38) could have formulated the problem I solved.

## II. Substance

Consistency conditions, like any equilibrium condition with a rational expectations flavor, may well be too strong to describe real-world oligopolies. But Daniel errs in his presumption that relaxing consistency leads inevitably to Cournot (quantity) equilibrium.

On the contrary, it leads to the completely indeterminate oligopoly problem. Daniel's defense of Cournot would ring as true if everywhere the word "Cournot" were replaced by "Bertrand"; "quantity" by "price." This is the core issue of noncooperative oligopoly theory; competition could be formalized by any of several games, each perfectly sensible mathematically. The point of consistency conditions is to remove this indeterminacy.

Daniel's further allegation, that imperfect information somehow leads to Cournot, is also incorrect. In some recent work in which imperfect information is explicitly treated, Arthur Robson and Stephen Turnbull find the *CCE* as the correct equilibrium.

Daniel's allusions to the theory of limit pricing/entry deterrence and to perfect competition strike me as odd. Limit pricing raises exactly the same issues of solution concept as does duopoly. The Cournot theory of entry deterrence has been widely studied under the label "Bain-Sylos Labini Postulate." The Bertrand theory has recently received much attention under the rubric of "sustainability." Perfect competition, on the other hand, raises no such issues; this is merely the point that atomistic firms need not consider their effect on prices.

## III. Conclusion

It is inevitable that oligopolists be concerned about one another's actions, and that the mathematical theory of oligopoly have some indeterminacy. I have recently proposed both empirical and theoretical approaches to the resolution of this indeterminacy. The *CCE* picks one solution concept for noncooperative oligopoly on theoretical grounds, leaving to empirical research the problem of telling competition from collusion. This strikes me as one sensible division of labor.

\*Assistant professor of economics, Stanford University.

## REFERENCES

- Bowley, A. L., *The Mathematical Groundwork of Economics*, Oxford: Oxford University Press, 1924.
- Bresnahan, Timothy F., (1981a) "Departures from Marginal-Cost Pricing in the American Automobile Industry: Estimates for 1977 -78," *Journal of Econometrics*, November 1981, 17, 201-27.
- \_\_\_\_\_, (1981b) "Duopoly Models with Consistent Conjectures," *American Economic Review*, December 1981, 71, 934-45.
- \_\_\_\_\_, "The Oligopoly Solution Concept is Identified," *Economics Letters*, November 1982, 10, 87-92.
- Daniel, Coldwell III, "Duopoly Models with Consistent Conjectures: Comment," *American Economic Review*, March 1983, 73, 238-39.
- Robson, Arthur J., "Implicit Oligopolistic Collusion is Destroyed by Uncertainty," *Economics Letters*, March 1981, 7, 144-48.
- Turnbull, Steven J., "Duopoly Models with Consistent Conjectures: Comment," mimeo., Stanford University, 1982.

# Theory of Screening and the Behavior of the Firm: Comment

By SHIRO YABUSHITA\*

In his article on the theory of screening in this *Review*, Joseph Stiglitz (1975) examines the characteristics of the equilibrium under imperfect information about workers' abilities. In the simple two-type model, in contrast to A. Michael Spence's signalling model (1973, 1974) where high quality workers have a lower marginal cost of signalling activity, Stiglitz introduces a screening process where the more able workers can take a perfectly accurate test at a fixed cost. The firm in his model seems to be limited in its activity so that it is not so responsive to profitable possibilities. The concept of equilibrium differs depending upon the assumptions concerning the behavior of the firm as Michael Rothschild and Stiglitz, and others point out.<sup>1</sup>

In this comment, I reconsider the Stiglitz screening model, assuming that the behavior of the firm is more responsive to profitable possibilities than that considered by Stiglitz. I show that the screening process in the Stiglitz model has similarities with that in the Spence model in some aspects. I also point out other characteristics of the Stiglitz model, considering the existence of equilibria in the cases of no social return from screening and of positive social return from screening, and show that the existence of equilibria depends upon the supposed behavior of the firm and the concept of equilibrium in the Stiglitz model.

## I

Let us consider the first case, no social return from screening, as analyzed by Stig-

litz. The productivities of the more able and the less able workers are denoted by  $\theta_1$  and  $\theta_2$ , respectively, and the mean productivity of the population is  $\bar{\theta} = \theta_1 h + \theta_2(1 - h)$ , where  $h$  is the fraction of the more able workers and  $\theta_1 > \bar{\theta} > \theta_2$ . Free entry and perfect competition will ensure that wage policies chosen by the firm in market equilibrium make zero expected profits. Necessary conditions for the existence of no-screening and full-screening equilibria are derived as follows. If the productivity of the more able workers is higher than the sum of the mean productivity and the screening cost per worker  $c$ , that is,  $\theta_1 > \bar{\theta} + c$ , the no-screening equilibrium does not exist. On the other hand, if  $\theta_1 < c + \theta_2$ , the full-screening equilibrium does not exist. Stiglitz concludes that under assumption (1) (herein reproduced):

$$(1) \quad \theta_1 - \theta_2 > c > \theta_1 - \bar{\theta},$$

there may be multiple equilibria.

In the Stiglitz model, it has been assumed that the wages paid to the workers are equal to their respective productivities if workers' abilities are known and to the mean productivity if unknown, and that the firm does not like any other wage policies even if they may be profitable. Under these assumptions, if  $\theta_1 - \bar{\theta} \leq c \leq \theta_1 - \theta_2$ , there may be both the no-screening and full-screening equilibria. However, the above results will not be true if we suppose that the firm may adopt other wage policies if they are more profitable than the above payment system.

Let us consider the full-screening equilibrium in the case where condition (1) holds. In the full-screening equilibrium, the more able worker receives a net income of  $\theta_1 - c$  and the less able an income of  $\theta_2$ . In this case, the firm can make a profit, without any information about abilities of workers, from paying such a wage,  $w$ , that is higher than the productivity of the more able net of the

\*Yokohama National University. I acknowledge the suggestions and comments of J. E. Stiglitz, M. Okuno, and the referee. My research was financially supported by the Seimeikai and the Tokyo Center for Economic Research.

<sup>1</sup>The behavior of the firm supposed by Stiglitz is different from that of the insurance company supposed by Rothschild and Stiglitz; the latter is more responsive than the former. I shall suppose the latter type of the firm in the later arguments.

screening cost, that is,  $\bar{\theta} > w > \theta_1 - c > \theta_2$ . Under this wage system both the more able and the less able workers can receive higher wages than in the full-screening equilibrium, but the total wage bills do not exceed the total amount of output. Since both the firm and workers prefer this wage offer, the full-screening equilibrium above does not exist as long as  $\theta_1 - \bar{\theta} < c$ . The above argument about the full-screening equilibrium is shown in Figure 1, where the wage and productivity of the more able workers are measured horizontally and those of the less able, vertically. Point *A* corresponds to the productivities of the two types of workers, that is,  $\theta_1$  and  $\theta_2$ . A pair of net incomes for the two types of workers in the full-screening equilibrium is given by point *D* under assumption (1) where the screening cost is indicated by the distance *AD*. On the other hand, their incomes in the no-screening equilibrium are  $\bar{\theta}$  as indicated by point *B*. In the neighborhood of *B*, there exist points such as *E* on the 45° line which are preferable to *D* for both the firm and workers so that the full-screening equilibrium does not exist.

Thus far, I have examined the full-screening equilibrium with the wages for the two types of workers being equal to their respective productivities. For full-screening, however, the firm may choose wages for the more able and the less able workers,  $w_1$  and  $w_2$ , so that they satisfy the following zero-profit condition:

$$w_1 h + w_2 (1 - h) = \bar{\theta} = \theta_1 h + \theta_2 (1 - h)$$

and so that the net income of the more able is more than the wage for the less able:  $w_1 - c > w_2$ . These wage policies are indicated by points on the line *CS* in Figure 1, and they imply subsidies from one type of workers to the other. But such full-screening equilibria with subsidies between the two types of workers cannot exist. This can be shown as follows; let us consider such a wage policy as *A'*. If the firm offers another pair of wages like *A''*, the less able workers prefer *A''* to *A'*, but the more able do not. The wage policy of *A''* is profitable for  $w_2 < \theta_2$ , but the firm loses money at *A'* so that the full-screening equilibrium at *A'* is not viable. For the

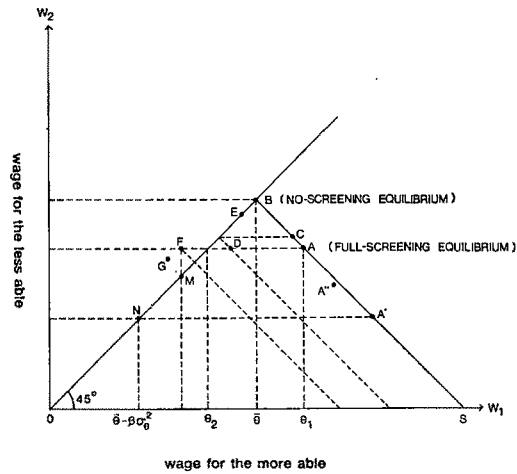


FIGURE 1

same reason, the equilibrium with subsidies from the more able to the less able cannot exist.

The above argument is exactly the same as that made by John Riley (1975) against the Spence model. The existence of the full-screening equilibrium results from the behavior of the firm assumed in the Stiglitz model just as the existence of the signalling equilibria depends on the behavior of the firm in the Spence model.<sup>2</sup> Contrary to Riley's argument, however, the no-screening equilibrium exists under assumption (1). Moreover, in the Stiglitz screening model with a single accurate test of fixed cost, there is a unique equilibrium, except in the borderline case of  $\theta_1 - \bar{\theta} = c$ , where both the full-screening and no-screening equilibria may exist. As long as there are multiple equilibria, Stiglitz's conclusion about the Pareto inferiority of the full-screening equilibrium still holds.<sup>3</sup>

## II

Next let us consider the case of information externalities also analyzed by Stiglitz. It is assumed that the output per worker of the

<sup>2</sup>Although Stiglitz argues that multiple equilibria occur in his model for different reasons than in the Spence model, it is not necessarily so.

<sup>3</sup>The more able worker earns the same net income  $\theta_1 - c = \bar{\theta}$  in the two equilibria, while the income of the less able is  $\bar{\theta}$  in the no-screening equilibrium and  $\theta_2$  in the full-screening one.

assembly line  $\bar{\theta} - \beta\sigma_\theta^2$  is lower than the productivity of the less able worker, and that a single individual cannot raise his income by buying screening. Stiglitz shows that under assumption (6) (herein reproduced):

$$(6) \quad \theta_1 - \bar{\theta} < \theta_1 - \theta_2 < c < \theta_1 - \bar{\theta} + \beta\sigma_\theta^2,$$

the full-screening equilibrium does not exist and that the market equilibrium involves no screening. For the more able worker does not buy screening if the screening cost per worker is such that  $\theta_1 - \theta_2 < c$ , where individuals receive an income of  $\bar{\theta} - \beta\sigma_\theta^2$  in the no-screening equilibrium.<sup>4</sup> That is, even if the more able worker screens himself from the less able, his wage does not rise since the firm does not respond to the screening. In such a case, however, the firm may react if it is profitable.

Does the no-screening equilibrium still exist if the firm reacts in such a way? Let us suppose that the firm is more responsive to the screening, that is, the firm assigns the screened workers to a separate assembly line if it is profitable. Then the output per screened worker will be raised from  $\bar{\theta} - \beta\sigma_\theta^2$  to  $\theta_1$  and the firm can raise the wage rate for the screened worker up to  $\theta_1$ . Thus, if the screening cost per worker is such that  $c < \theta_1 - \bar{\theta} + \beta\sigma_\theta^2$ , the more able worker will buy screening and the no-screening equilibrium does not exist.<sup>5</sup> We can represent this in Figure 1 as follows. Suppose Point *N* is the no-screening equilibrium and *F* indicates a pair of net incomes for the two types of workers in the full-screening equilibrium. It is clear that *F* is preferable to *N* for both the

more able and the less able workers. If the firm offers different wages for the workers, which correspond to a point like *G* near *F*, as described above, it can make a profit and all workers will be better off so that the no-screening equilibrium cannot exist. Thus, in this case, contrary to the results derived by Stiglitz, neither the full-screening nor the no-screening equilibrium exist under assumption (6). The screening cost must be higher than or equal to  $\theta_1 - \bar{\theta} + \beta\sigma_\theta^2$  for the existence of the no-screening equilibrium.

The above nonexistence of the equilibria is pointed out by Riley (1975) in the signalling model.<sup>6</sup> Here we have the same result in the Stiglitz model as Riley obtains in the Spence model, though the two models assume different screening processes. As indicated in the previous case, the existence of the equilibria depends upon the assumed behavior of the firm rather than the screening processes.

Charles Wilson (1977) and Riley (1979) have introduced alternative equilibrium concepts, taking into consideration the effects of reaction by other firms. In his insurance market model, with its alternative concept of equilibrium, Wilson shows that equilibria always exist and that the equilibrium becomes the no-screening equilibrium if the full-screening equilibrium does not exist. But in the case of positive social return from screening, no equilibria exist under Stiglitz's assumption (6) even in Wilson's sense of the word.<sup>7</sup> In this model, the full-screening equilibrium requires  $w_1 - c > w_2$ , but full-screening equilibria with subsidies among workers cannot exist as shown in the previous section. Such full-screening equilibria cannot be the Wilson equilibrium. For the wage policy of *A''* is still profitable to the firm after the unprofitable offer of *A'* is withdrawn.

On the other hand, in the no-screening equilibrium, the firm pays both the more able and the less able workers the same income of  $\bar{\theta} - \beta\sigma_\theta^2$ , which is the per worker productivity at the assembly line. After the

<sup>4</sup>In this case, the existence of full-screening equilibrium does not require  $c < \theta_1 - \bar{\theta}$  as in the preceding case, but  $c < \theta_1 - \theta_2$  as in the argument given by Stiglitz. Since in the no-screening equilibrium the firm cannot offer wages which are higher than those in the full screening, the nonexistence of full-screening equilibrium results only from the myopic action of the more able workers caused by their net income being less than the income of the less able.

<sup>5</sup>If the firm reacts in such a way, the less able worker may have incentive to be screened since it is assumed that the output per worker of the assembly line is less than the productivity of the less able worker. He may receive higher income from screening, but does not have any screening process here.

<sup>6</sup>Rothschild-Stiglitz (1976) and Charles Wilson (1977) point out the same problem in their model of insurance markets.

<sup>7</sup>In this sense the Stiglitz model differs from the Spence model.

more able workers are attracted to a separate assembly line as discussed above, only the less able are left at the old assembly line so that its productivity becomes  $\theta_2$  instead of  $\bar{\theta} - \beta\sigma_\theta^2$ . From (6), the productivity of the workers exceeds the wage rate, that is,  $\theta_2 > \bar{\theta} - \beta\sigma_\theta^2$  at the old line so that the line continues to operate. Thus, the no-screening equilibrium does not exist even in Wilson's concept.

Finally, I comment on the existence of the partial-screening equilibrium; an equilibrium involving partial screening does exist in the case of positive return from screening if the wage or per capita productivity on the unscreened assembly line is a continuous function of the fraction of the more able workers who are screened.<sup>8</sup> Denote this fraction by  $s$  and let the wage on the unscreened assembly line be a function of it,  $\hat{w}_2 = \hat{w}_2(s)$ , where  $0 \leq s \leq 1$ . The wage on the unscreened assembly line can be  $\theta_2$ , that is,  $\hat{w}_2(1) = \theta_2$  if all the more able are screened while  $\hat{w}_2(0) = \bar{\theta} - \beta\sigma_\theta^2$  if none of them are screened. On the other hand, the net income for the screened more able workers is  $\theta_1 - c$ . Since  $\theta_2 > \bar{\theta} - \beta\sigma_\theta^2$ , there exists an  $s^*$  such that  $\hat{w}_2(s^*) = \theta_1 - c$ . Thus, under assumption (6), there exists a partial-screening equilibrium characterized by  $s^*$  and  $\hat{w}_2(s^*)$ . The pair of net incomes for the workers in this equilibrium is indicated by Point  $M$  in Figure 1, where the wage for the screened worker is measured horizontally and that for the unscreened

worker vertically in this case. However, in this equilibrium the more able workers can receive the same income on the unscreened assembly line as their net income from screening. Thus, the fraction  $s^*$  of the more able do not necessarily but only coincidentally screen themselves under the wage system of  $\hat{w}_2(s^*) = \theta_1 - c$  as long as they randomly decide whether or not to buy screening. In this sense, the above partial-screening equilibrium is not stable.

## REFERENCES

- Riley, John G., "Competitive Signalling," *Journal of Economic Theory*, April 1975, 10, 174-86.
- , "Informational Equilibrium," *Econometrica*, March 1979, 47, 331-59.
- Rothschild, Michael and Stiglitz, Joseph, "Equilibrium in Competitive Insurance Markets: The Economics of Imperfect Information," *Quarterly Journal of Economics*, November 1976, 90, 629-49.
- Spence, A. Michael, "Job Market Signalling," *Quarterly Journal of Economics*, August 1973, 87, 355-79.
- , "Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution," *Journal of Economic Theory*, March 1974, 7, 296-332.
- Stiglitz, Joseph E., "The Theory of 'Screening,' Education, and the Distribution of Income," *American Economic Review*, June 1975, 65, 286-300.
- Wilson, Charles, "A Model of Insurance Markets with Incomplete Information," *Journal of Economic Theory*, December 1977, 16, 167-207.

<sup>8</sup>The existence of this partial-screening equilibrium is pointed out by Stiglitz.

# Alternative Approaches to Analyzing Markets with Asymmetric Information: Reply

By J. STIGLITZ AND A. WEISS\*

Shiro Yabushita's comment on "The Theory of Screening" (Stiglitz, 1975) raises some interesting questions which have arisen in a number of different contexts: How should one model markets with imperfect information? What are appropriate behavioral postulates and equilibrium concepts? To restructure the question in terms of game theoretic language: Who are the active players? What are their strategy spaces? Do the players move simultaneously, or does one move before the other?

These questions do not have simple answers: different equilibrium notions are appropriate for different economic contexts. Yabushita's comment illustrates how one can obtain markedly different results using alternative assumptions. Although we believe that for the particular context examined in Stiglitz (1975), the equilibrium concept employed there is better than the alternative implicitly proposed by Yabushita, his formulation merits attention, and there are other contexts in which his would be the appropriate formulation.

In order to illuminate how the Yabushita model differs from the Stiglitz model, it is useful to reformulate their analyses in game theoretic terms. The strategies of individuals entail decisions as to whether or not to screen themselves (go to school), while the strategies of firms are wage offers conditioned on whether or not the individual has screened himself, and the result of that screening. An individual goes to work for the firm which offers him the highest wage. Thus, the payoff to the firm is zero, if it doesn't obtain the worker, and the difference between the worker's productivity and his wage if it obtains the worker. The payoff to the individ-

ual is his maximum wage offer minus his screening costs.

There are two alternative ways of viewing the difference between the Yabushita and Stiglitz formulations of the education market. One way is to view Yabushita and Stiglitz as differing in their assumptions about whether firms or individuals move first.

The choice of a schooling program is typically made by individuals *before* a wage offer is received. The Stiglitz model attempts to capture this natural temporal sequence of decision making by formulating a two-period model; in the first period, individuals decide whether or not to screen themselves, while in the second, profit-maximizing firms offer wages to both screened and unscreened individuals (who, in turn, go to work for the firm offering them the highest wage). The wage offer of each firm is assumed to be the profit-maximizing response to the screening decision of workers, given the wage offers of other firms. Hence, if any high ability worker screens himself he will receive a wage equal to his expected productivity  $\theta_1$ .<sup>1</sup> If no workers screen themselves, the unscreened workers receive a wage equal to their expected productivity of  $\bar{\theta}$ , and if the type 1 workers screen themselves, the unscreened workers receive a wage of  $\theta_2$ , their expected productivity. In each case, higher wages would incur losses; and a lower wage could not be an equilibrium, since there would be an opportunity for some other firm to increase its profits by offering a higher wage.<sup>2</sup>

<sup>1</sup>This is true whether or not any other workers are screening themselves. Consequently, we are precluding Nash equilibria in which the strategies of all firms are to offer low wages to screened workers and the strategy of all workers is to not screen themselves. Rather, we are assuming that once a screening decision is made, firms always respond optimally at that point. See fn. 4 below.

<sup>2</sup>That is, if each firm takes the wage offers and screening decisions as given, unless the maximum wage

\*Princeton University and Bell Laboratories, respectively.

In this interpretation of Stiglitz, if no one is screening himself, a high ability type would only increase his income by screening if his productivity, net of screening costs  $c$ , exceeds the productivity of a *randomly selected individual* in the population: those are the relevant wages with and without screening starting from a position where no one screens himself. Thus, if  $\bar{\theta} > \theta_1 - c$ , an equilibrium exists with no screening.

At the same time, a screening equilibrium would also exist if the productivity of the most able, net of screening costs, exceeds the productivity of the *least able*. These provide the relevant wages in a screening equilibrium. Obviously, in this formulation, multiple equilibria may exist. Notice that we are implicitly assuming that there are many workers of each type so that the choice of a single individual has trivial effects on the wage offers of firms. Implicitly we are defining equilibrium as satisfying a backward induction argument, so that workers know that for any choice of an education level (screen), every firm responds in a profit-maximizing way given the strategies of all other firms and workers. This formulation is close to the notion of subgame-perfect equilibrium. (A subgame-perfect equilibrium is a Nash equilibrium which, if written in extensive form, also generates a Nash equilibrium for any game beginning at an information set of the original game containing only one node.)

On the other hand, in the Yabushita analysis, firms make their wage offers prior to individuals choosing whether or not to screen themselves, and the screening decisions of individuals are the optimal reactions to these wage offers of firms. Given this structure, multiple equilibria cannot exist. If the conditions for a pooling equilibrium are satisfied, that is,  $\bar{\theta} > \theta_1 - c$ , any putative full-screening equilibrium would be broken by a firm offering a wage (conditional on no screening) of  $\bar{\theta} - \epsilon$  where  $\epsilon$  is less than  $\bar{\theta} - \theta_1 + c$ . The reason for nonexistence of multiple equilibria is that if firms are moving first, the

wage offer of any firm can determine whether or not individuals screen themselves.<sup>3</sup>

The Yabushita analysis of equilibrium in a market in which firms move first is similar in many respects to a Nash equilibrium in which individuals are not active players in the game. This is the second interpretation for the difference between Stiglitz and Yabushita: Stiglitz is looking at Nash equilibria in which both firms and individuals are active players; Yabushita is looking at Nash equilibria in which only firms are active players. That is, individuals simply work for the firm offering the best contract to them. They do not choose strategies, but simply react to the strategies of the firms. When a firm decides on a strategy, it takes the reaction functions of individuals as given.

This assumption of passive individuals is a plausible one to make if the screening cost  $c$  is some preemployment exam administered by the firm as in Guasch and Weiss (1980). Similarly, in the insurance market studied by Rothschild-Stiglitz (1976), it is plausible to think of individuals as passively choosing from a set of insurance policies. However, in the Stiglitz (1975) education model where the screen is implicitly a course of schooling with the property that the more able pass, and the less able fail, it seems implausible to assume that individuals are simply reacting to the strategies of firms. It seems more sensible to view that schooling program as a strategic choice of an individual. Both individuals and firms are active players in this market. Individuals decide whether or not to screen themselves (go to school) and firms make

---

offer equals the expected productivity, the firm can increase its profits by offering an amount slightly over that offered by other firms.

<sup>3</sup>Some readers may be puzzled why, in such a situation there is a pooling equilibrium: Michael Rothschild and Stiglitz (1976) and Charles Wilson (1980) show that, in general, there cannot exist a pooling equilibrium. The reason is that here we assume that there is a fixed cost,  $c$ , associated with identifying oneself as more able and we have not allowed the firm to employ alternative self-selection devices. If, for instance, the firm can require individuals to disclose whether they are more or less able, and can insist that individuals post a bond, which they forfeit if, upon subsequent screening, it turns out that they have lied, then the pooling equilibrium can be broken by a self-selection equilibrium (which itself may be broken by a pooling equilibrium) entailing random testing; under these circumstances, equilibrium may not exist. (See Stiglitz, forthcoming; David Scharfstein, 1982; and J. Luis Guasch and Weiss, 1981.)

wage offers conditional on whether or not the individual has gone to school. With this interpretation, the assumption that the output of a randomly selected individual exceeds the productivity of the most able worker, net of screening costs, which in turn exceeds the output of the least able worker (i.e.,  $\theta_1 - \theta_2 > c > \theta_1 - \bar{\theta}$ ) does indeed imply multiple equilibria.<sup>4</sup> The cost  $c$  of the screen is a sunk cost. In a full-screening equilibrium, the more able are paid a wage equal to the value of their output and would not be attracted to a firm offering a lower wage. Their lifetime income is  $\theta_1 - c$ . They may prefer a wage equal to the expected productivity of a randomly selected individual, but that is just to say that no one likes Pareto-dominated equilibria, not that those equilibria do not exist. Weiss (forthcoming) shows that with more realistic assumptions such as a continuum of possible education levels, many types of individuals, imprecise estimates by each worker of his own ability, and imprecise tests, the problems of multiple equilibria do not disappear. Some of the Nash equilibria which arise are clearly implausible, and Weiss introduces more restrictive definitions of equilibrium to eliminate those implausible equilibria.

Section II of Yabushita's comment is motivated by the different implications of signals purchased by the workers and tests administered by the firms. In that section, the difficulty lies in understanding how firms respond to a worker who has screened himself. Because of the assumption of assembly line production, no single worker would gain from screening himself. On the other hand, a coalition of able workers would benefit from screening themselves (and the less able would

even find it profitable to subsidize this screening). Yabushita again assumes either that firms do the screening as in Guasch and Weiss (1980), or that individuals passively respond to the wage offers of firms in deciding whether or not to screen themselves. Thus firms do not take the screening decisions of individuals as given when choosing their wage offers; rather, firms make their wage offers before workers screen themselves, and the latter do not act strategically. The lesson to be drawn from these examples is that in adverse selection models it is important to correctly delineate who is actively signaling or screening.

Elsewhere, we have reexamined the nature of the equilibrium which emerges in a variety of market situations. There are some situations, such as the insurance market or labor markets where firms engage in some screening, where the uninformed are active players, while the informed are passive; in those situations the kinds of considerations which Yabushita has raised are central (paralleling the earlier results of Rothschild and Stiglitz). There are other situations, such as the education market examined in Stiglitz (1975), A. Michael Spence (1974), and Weiss (1982), or the capital market, as examined in Stiglitz (1982) and Sudipto Bhattacharya (1980), where the informed are active players, and move prior to the uninformed.

The recent literature on equilibrium with imperfect and costly information has established that the simple price-taking formulations associated with Walrasian general equilibrium analysis may well be inappropriate in such situations.<sup>5</sup> What we have attempted to show in this response to Yabushita's insightful comment is that the choice of an appropriate equilibrium concept entails a careful analysis of the economic structure of the problem at hand. There are no easy or general answers.

<sup>4</sup>If we do not restrict the strategy space of firms to pay types who have screened themselves wages equal to their expected productivities, the only requirement for multiple equilibria is that  $\theta_1 - c > \theta_2$ : there are always no-screening Nash equilibria. This follows from the definition of Nash equilibrium which allows firms to react nonoptimally to out-of-equilibrium moves. For example, if no workers are screening themselves, firms could offer a zero wage to screened workers. In that case, no screening would be an equilibrium regardless of the value of  $\theta_1$ .

<sup>5</sup>For instance, while traditional Walrasian analysis defines equilibrium as having zero excess demands, Weiss (1980), Stiglitz and Weiss (1981), Stiglitz (1976) and Wilson (1980) have shown that competitive equilibrium may entail excess supply (of labor) or excess demand (for loans).

## REFERENCES

- Akerlof, George, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, August 1970, 84, 448-500.
- Bhattacharya, Sudipto, "Non-Dissipative Signaling Structures and Dividend Policy," *Quarterly Journal of Economics*, August 1980, 95, 1-24.
- Guasch, J. Luis and Weiss, Andrew, "Wages as Sorting Mechanisms in Markets with Asymmetric Information: A Theory of Testing," *Review of Economic Studies*, July 1980, 48, 653-64.
- \_\_\_\_\_, and \_\_\_\_\_, "A Note on Self-Selection in the Labor Market: Random Testing Revisited," Bell Laboratories Discussion Paper No. 230, 1981.
- Rothschild, Michael and Stiglitz, Joseph, "Equilibrium in Competitive Insurance Markets," *Quarterly Journal of Economics*, November 1976, 90, 629-49.
- Scharfstein, David, "On Inducing Honest Behavior: Towards a General Theory of Self-Selection and the Design of Predation Policy," senior thesis, Princeton University, 1982.
- Spence, A. Michael, "Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution," *Journal of Economic Theory*, March 1974, 7, 296-332.
- Stiglitz, Joseph, "The Theory of 'Screening,' Education and the Distribution of Income," *American Economic Review*, June 1975, 65, 283-300.
- \_\_\_\_\_, "Prices and Queues as Screening Devices," IMSSS Technical Report, 1976.
- \_\_\_\_\_, "Information and Capital Markets," in W. S. Sharpe and C. M. Cootner, eds., *Financial Economics: Essays in Honor of Paul Cootner*, Englewood Cliffs: Prentice-Hall, 1982, 118-58.
- \_\_\_\_\_, *Information and Economic Analysis*, Oxford: Oxford University Press, forthcoming.
- \_\_\_\_\_, and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, 71, 393-409.
- Weiss, Andrew, "Job Queues and Layoffs in Markets with Flexible Wages," *Journal of Political Economy*, May 1980, 88, 526-58.
- \_\_\_\_\_, "A Sorting-cum-Learning Model of Education," *Journal of Political Economy*, forthcoming.
- Wilson, Charles, "The Nature of Equilibrium in Markets and Adverse Selection," *Bell Journal of Economics*, Spring 1980, 11, 108-30.
- Yabushita, Shiro, "Theory of Screening and the Behavior of the Firm: Comment," *American Economic Review*, March 1983, 73, 242-45.

## ERRATA

### Product Differentiation Advantages of Pioneering Brands

By RICHARD SCHMALENSEE\*

I am indebted to Robert Porter of the University of Minnesota for pointing out a number of errors in my paper published in the June 1982 issue of this *Review*. The conclusions of that paper are in no way affected by the changes that follow.

There is a typographical error in the second line of the second paragraph of Appendix A, p. 361:  $\Pi(p^0)$  should be  $\Pi^0(p^0)$ . Appendix B, pp. 362-63, contains a number of algebraic and typographical errors. A correct execution of the analysis outlined there supports the following revision of p. 356 of the text, beginning with the second complete sentence in the second column:

...It can be shown that if  $r < 1$  and  $\tau \leq 1/3$ , the second brand optimally charges  $(P_1 - \tau V)$  in the first period and (just under)  $P_1$  thereafter. This involves moving to a point like *A* in Figure 3 and selling to all who had previously purchased the pioneer. The net present values of the two brands are related by

$$(6) \quad W_2 = W_1 - \tau^2/4(1 - \tau) \\ = (1/4) \left[ \frac{1 - 2\tau}{1 - \tau} + \frac{1}{r} \right] - F.$$

If  $\tau = 0$ ,  $W_2 = W_1$ ; there is no quality uncertainty and there is obviously no barrier. If  $F = 0$ , so that there are no scale economies,  $W_2$  is always positive but less than  $W_1$ . The second brand...

\*Massachusetts Institute of Technology.

## ERRATA

### Inventory Investment and the Theory of the Firm

By LANNY ARVAN AND LEON N. MOSES\*

The optimal policy reported for Case III in our article in the March 1982 issue of this *Review* is incorrect. The correct optimal policy has a production path that is discontinuous everywhere, that is, production is on, off, on, off, etc. When on, the rate of production is that at which average cost is a minimum. At almost every time, inventory is zero so that the storage cost function also takes on the value zero. Storage capability is utilized to transfer product from an on-production time to an off-production time. The firm does not price discriminate. At each point of time it sells that quantity at which marginal revenue equals minimum average cost.

Positive inventory accumulation will take place in a continuous time model if the economies of scale are in the production of a stock of output rather than in the production of flows as in the paper. This condition naturally arises when there are start-up costs. In effect, then, output units are measured differently than sales units. An alternative way of stating this is that to exhaust any positive output requires sales over some time interval. When output is measured in stocks and the average cost function is U shaped, inventory will be accumulated, sales will be as described in the paper, but production will occur at distinct points in time rather than over an interval of time.

If the model is cast in discrete time, there is no need to measure output and sales in different units. The optimal policy is then as described above.

\*University of Illinois at Urbana-Champaign and Northwestern University, respectively.

## NOTES

The Economics Faculty at the University of Colorado wish to express their concern over the disappearance and presumed death of their colleague, Nicholas W. Schrock. Professor Schrock disappeared on May 31, 1982, while driving between Hemosilla and Mazatlan, Mexico, on his way to a summer teaching position in Guadalajara. A trial of seven policemen on charges relating to his disappearance began on July 30, but as yet no findings have been released. Contributions to reimburse Mrs. Schrock for investigatory expenses may be sent to Globe Industrial Bank, P.O. Box 1069, Boulder, CO 80306. Contributions to a scholarship fund in honor of Professor Schrock may be sent to the Schrock Scholarship Fund, University of Colorado Foundation, Inc., University of Colorado, Boulder, CO 80309.

The International Economic Association is organizing in 1983 various roundtable conferences leading up to a world congress on "Structural Change, Economic Interdependence and World Development." Authors of recent contributions to aspects of the above subject matter are invited to send abstracts of their work to the Secretariat of the IEA, 4 rue de Chevreuse, 75006 Paris, France, so that they may be forwarded to the different organizers likely to be interested in their contributions of a roundtable conference or for the Congress. A copy might also be sent to the President of the IEA, Professor Victor L. Urquidí, El Colegio de México, Camino al Ajusco No 20, México 20, D.D., Apartado Postal 20-671, México.

*New Journal:* Spring 1983 is the first quarterly issue of the *Journal of Marine Resource Economics*, edited by Jon G. Sutinen, University of Rhode Island. The editorial board includes: James A. Crutchfield, University of Washington; Brian Rothschild, University of Maryland; Francis T. Christy, Jr., FAO, Rome; Fred Prochaska, University of Florida; Michael Sissenwine, Northeast Fisheries Center; and James K. Sebenius, Harvard University.

The Arne Ryde Symposium on the Primary Sector in Economic Development is being arranged by the Department of Economics, University of Lund, August 29-30, 1983. The topic has been chosen against the background of the importance of agriculture and mineral resources in the development of the Third World. Abstracts should be sent by April 15, 1983. Participants will receive free board and room at the Frostavallen Conference Hotel. Those presenting papers will have their travel expenses covered. (For participants from outside Scandinavia, however, it may not be possible to reimburse costs in full.) Address all correspondence to

the Secretary: Åsa Weibull, Nationalekonomiska institutionen, Magistratsvägen 55 N, S-222 44 LUND, Sweden. (Telephone: 046-10 86 82.)

*Call for Papers:* The thirteenth Annual Conference of the Illinois Economic Association will be held on October 28-29, 1983, in Chicago. Proposals for papers for presentation should be submitted by May 1, 1983, to Dr. John Mathis, Economics Research Division, Continental Illinois Bank, 231 S. LaSalle Street, Chicago, IL 60693.

The *Monograph Series in Economics and Finance* is published by the Salomon Brothers Center for the Study of Financial Institutions at the Graduate School of Business Administration at New York University. The series publishes four monographs per year on a broad range of subjects: Manuscript submissions are invited and should be longer than the typical journal article, but shorter than most books. Monographs can present reports on scholarly research or be of a descriptive nature. A rapid review is promised. Submissions should be sent in duplicate to the editors, Ernest Bloch and Paul Wachtel, New York University Graduate School of Business Administration, 90 Trinity Place, New York, NY 10006.

The International Symposium of Forecasting will meet in Philadelphia, June 5-8, 1983. The conference focuses on research on forecasting methods and on the use of forecasts. It is sponsored by the International Institute of Forecasters in collaboration with the Wharton School, University of Pennsylvania. Send abstracts of 100 words or less to Professor J. Scott Armstrong, Wharton School, University of Pennsylvania, Philadelphia, PA 19104. (Telephone 215 + 898-5087.)

*Call for Papers:* The 4th volume of *Research in Human Capital and Development* (an annual compilation of research published by JAI Press Inc.) will be titled *Migration Theory, Human Capital and Development*. Manuscripts are invited that make a significant contribution to the topic. Manuscripts will be refereed and contributors will receive royalties on a page-weighted basis. Send three copies with abstracts to either of the editors: Ismail Sirageldin, Departments of Population Dynamics and Political Economy, The Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, or Oded Stark, Department of Economics, Harvard University, Cambridge, MA 02138.

The annual meeting of the Population Association of America will be held April 14-16 in Pittsburgh, Pennsylvania. Sessions on demographic aspects of labor markets and economic development are included in the program. For a preliminary program and registration materials, contact John L. Goodman, Jr., The Urban Institute, 2100 M Street NW, Washington, D.C. 20037.

The Association of Environmental and Resource Economists (AERE) requests nominations for two awards in the fields of resource and environmental economics. The first award will be for the best published work in the field by a scholar under age 35, and will be made annually. It is anticipated that senior professionals, such as the candidate's dissertation supervisor, will propose the work to either V. Kerry Smith, Chair of the Awards Committee, or for nominations outside the United States, to Geoffrey M. Heal. The nomination should include a statement describing the contribution of the paper, plus two copies (or reprints) of the paper. The second award will be made for contributions of enduring quality to the field over a period of ten years or more. This award will be made as appropriate. Two copies of a 2-4 typed-page description of the contribution should be submitted. The Awards Committee consists of A. C. Fisher, G. M. Heal, and V. K. Smith. The deadline for nominations for both awards is July 1, 1983. Winners will be announced at the 1983 AERE annual meetings. For more information, contact V. Kerry Smith, Department of Economics, University of North Carolina, Gardner Hall 017A, Chapel Hill, NC 27514, or Geoffrey M. Heal, Department of Economics, University of Essex, Wivenhoe Park, Colchester, England CO4 3SQ.

The Emory University Law and Economics Center will hold a Legal Institute for Economists, June 19-July 1, 1983 at the Hanover Inn and Dartmouth Lodge dormitory, New Hampshire. Director Henry G. Manne heads the two-week course which provides an intensive introduction to the American legal system, its institutions, research methods, and procedures. It is a program in law, not law and economics. For further information, contact Mr. Marc Hoberman, Program Administrator, Emory University, Law and Economics Center, Atlanta, GA 30322. (Telephone 404 + 329-5768.)

Fordham University is sponsoring a two-day symposium on Ridge Regression Methods and Application, at the Lincoln Center campus, July 26-27, 1983. For further information, contact Professor H. D. Vinod, Economics Department, Fordham University, Bronx, NY 10458.

The tenth annual conference of the European Association for Research in Industrial Economics will be held

August 23-25, 1983 at the Norwegian School of Economics and Business Administration, Bergen, Norway. Sessions include Economics of Organization of Industry of Firms; Industrial and Competition Policy, Industrial Organization and Trade; Industry Studies; Regulation of Industries and Firms; Technological and Organizational Change and Industry Structure, as well as other issues reflecting current research in Industrial Economics. One-page abstracts should be sent immediately to the Chairman of the Programme Committee: Dr. Einar Hope, Director of Research, Center for Applied Research, Norwegian School of Economics and Business Administration, Helleveien 30, N-5035 Bergen-Sandviken.

The tenth annual meeting of the European Finance Association is scheduled for September 1-3 at the European Institute of Business Administration (INSEAD), Fontainebleau, France. Those who wish to present a paper should send a copy, or a detailed abstract, *immediately* to Professor Claude Viallet, INSEAD, Boulevard de Constance, F-77305 Fontainebleau Cedex, France.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

#### Deaths

Deane Carson, professor, Graduate School of Business, Columbia University, August 25, 1982.

Abba Lerner, Tallahassee, Florida, October 27, 1982.

M. G. Mueller, Department of Political Economy, University of Glasgow, Scotland, May 1982.

#### Retirements

Arthur G. Ashbrook, Jr., senior economist, Central Intelligence Agency, August 6, 1982.

Max R. Bloom, professor of real estate and urban land economics, Syracuse University, December 31, 1982.

#### Foreign Scholars

Andrzej Grochulski: visiting associate professor, department of economics, American University, 1982-83.

Francis Seton, Nuffield College, Oxford: professor of economics, The University of the South, January 1983-April 1983.

#### Promotions

Peter H. Calkins: associate professor of economics, Iowa State University, July 1, 1982.

Robert Deacon: professor of economics, University of California-Santa Barbara, July 1, 1982.

Philip Friedman: professor of economics, Boston University, September 1982.

Gilbert R. Ghez: professor of management and economics, Roosevelt University, August 15, 1982.

Kenneth Guentner: senior economist, Analysis Division, securities department, Federal Reserve Bank of New York, July 15, 1982.

Michael A. Hubbard: staff economist, U.S. Postal Service, Washington, D.C., June 26, 1982.

Wallace E. Huffman: professor of economics, Iowa State University, August 21, 1982.

Robert W. Jolly: associate professor of economics, Iowa State University, July 1, 1982.

Arthur E. King: associate professor of economics, Lehigh University, 1982.

Stephen A. McCafferty: associate professor of economics, Ohio State University.

Christopher J. McCurdy: research officer and senior economist, Open Market Operations Function, Federal Reserve Bank of New York, October 1, 1982.

Joseph P. Magaddino: professor of economics, California State University-Long Beach, August 30, 1982.

Marshall H. Medoff: professor of economics, California State University-Long Beach, August 30, 1982.

James W. Meehan, Jr.: professor of economics, Colby College, September 1982.

William H. Meyers: associate professor of economics, Iowa State University, July 1, 1982.

Cathy G. Miners: chief, Domestic Financial Markets Division, financial markets department, Federal Reserve Bank of New York, July 15, 1982.

Joe A. Stone: associate professor of economics, University of Oregon, September 15, 1982.

Wayne Thomas: professor, Agricultural Experiment Station, University of Alaska-Fairbanks, July 1982.

Jerry Thursby: associate professor of economics, Ohio State University.

Marie Thursby: associate professor of economics, Ohio State University.

Alden L. Toevs: associate professor of economics, University of Oregon, September 15, 1982.

#### Administrative Appointments

Evaldo A. Cabarrouy: director, graduate program in business administration, Universidad del Turabo, Puerto Rico, August 1982.

Philip Friedman: chairman, department of finance and economics, School of Management, Boston University, October 1982.

James W. Meehan, Jr.: chairman, department of economics, Colby College, September 1982.

Dennis M. O'Toole: associate dean for external affairs, School of Business, Virginia Commonwealth University, September 1, 1981.

Donald Phares: chairman, economics department, University of Missouri-St. Louis, August 1, 1982.

Andrew Postlewaite: chairman, department of economics, University of Pennsylvania, July 1, 1982.

Dominick Salvatore: chairman, department of economics, Fordham University, September 1, 1982.

#### Appointments

George Assaf, Swarthmore College: assistant professor of economics, American University, fall 1982.

Marianne Baxter: acting assistant professor of economics, University of California-Santa Barbara, July 1, 1982.

Valerie Bencivenga: acting assistant professor of economics, University of California-Santa Barbara, July 1, 1982.

Farrell E. Bloch, Econometric Research, Inc.: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, July 1982.

Edward Buffy: assistant professor of economics, University of Pennsylvania, July 1982.

Margriet Caswell: acting assistant professor of economics, University of California-Santa Barbara, July 1, 1982.

Ching Meei Lee Chang: assistant professor of economics, University of California-San Diego, July 1, 1982.

Walter A. Chudson, consulting economist: adjunct professor, School of Business Administration, University of Connecticut, September 1982.

Michael T. Devaney, University of Arkansas: assistant professor of finance, Memphis State University, September 1, 1982.

Edward J. Geng: senior vice president, Open Market Operations Function, Federal Reserve Bank of New York, September 28, 1982.

Robert R. Gottfried: assistant professor of economics, The University of the South, September 1982.

David Gray: assistant professor of economics, University of Pennsylvania, July 1982.

Kristin M. Hallberg, Amherst College: assistant professor of economics, Colby College, September 1982.

Kausar Hamdani: economist, Monetary Analysis Division, monetary research department, Federal Reserve Bank of New York, September 20, 1982.

Joni Hersch: assistant professor of economics, University of Oregon, September 15, 1982.

Daniel Himarios: instructor of economics, The University of the South, September 1982.

Richard L. Johnson, Civil Aeronautics Board: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, August 1982.

Willene Johnson: economist, Developing Economies Division, external financing department, Federal Reserve Bank of New York, September 13, 1982.

Martha Jones: economist, Developing Economies Division, external financing department, Federal Reserve Bank of New York, September 8, 1982.

Warren Jones: assistant professor of economics, School of Management, University of Alaska-Fairbanks, August 1981.

Herman I. Liebling: department of economics, Florida International University, August 1982.

Tsai-Fen Lin: assistant professor of economics, University of Cincinnati, September 1982.

Bonnie Loopesko: economist, Industrial Economies Division, international research department, Federal Reserve Bank of New York, August 9, 1982.

John G. Marcis: assistant professor of economics, Virginia Commonwealth University, August 16, 1982.

Fred B. Moseley, University of Massachusetts: visiting assistant professor of economics, Colby College, September 1982.

Samuel L. Myers, Federal Trade Commission: associate professor, Graduate School of Public and International Affairs, University of Pittsburgh, September 1, 1982.

Marc Nerlove: professor of economics, University of Pennsylvania, July 1982.

Randall J. Olsen, Yale University: associate professor of economics, Ohio State University.

William B. O'Neil, Amherst College: assistant professor of economics, Colby College, September 1982.

Mars A. Pertl, Southern Illinois University: professor of insurance, Memphis State University, September 1, 1982.

Pipat Pithyachariyakul, Northwestern University: assistant professor of economics, Rice University, July 1982.

Edward M. Rawson: adjunct assistant professor, department of economics, Iowa State University.

William J. Reid, Jr.: visiting assistant professor of economics, Virginia Commonwealth University, August 16, 1982.

David Sappington: assistant professor of economics, University of Pennsylvania, July 1982.

F. Bruce Simmons III, University of Cincinnati: as-

sistant professor of management and quantitative methods, University of Akron, 1982.

Marilyn J. Simon, Massachusetts Institute of Technology: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, August 1982.

Danny Steinberg, ABT Associates: department of economics, University of California-San Diego, July 1, 1982.

Judy Wachtenheim: economist, Business Conditions Division, domestic research department, Federal Reserve Bank of New York, September 15, 1982.

William B. Watkins, Middle Tennessee State University: assistant professor of finance, Memphis State University, September 1, 1982.

Douglas W. Webbink, Federal Communications Commission: Cornell, Pelcovits and Brenner Economists Inc., August 1982.

Kenneth J. White, Rice University: associate professor of economics, University of British Columbia, January 1982.

Nancy A. Williams: assistant professor of economics, School of Management, University of Alaska-Fairbanks, August 1982.

Douglas Woodham: economist, Industrial Economies Division, international research department, Federal Reserve Bank of New York, September 27, 1982.

#### Leaves for Special Appointments

M. Akbar Akhtar, Federal Reserve Bank of New York: Bank for International Settlements, Basel, Switzerland, November 1, 1982–October 30, 1983.

Sydney C. James, Iowa State University: Welfare Department, Church of Latter-Day Saints, Salt Lake City, September 1, 1982–July 31, 1983.

Edward M. Rawson, Iowa State University: Agency for International Development, Zambia.

Iva Lee Skov, California State University-Long Beach: Chief Administrator, Hall of Economics and Finance, California Museum of Science and Industry, August 30, 1982–83.

#### Resignations

Duane G. Harris, Iowa State University: General Mills, Inc., Minneapolis, Minnesota.

J. Stephen Henderson, Ohio State University: National Regulatory Research Institute, Columbus, Ohio.

Albert Keidel III, Ohio State University: Wharton Econometric Forecasting Associates, Washington, D.C.

William Manson, Ohio State University: Ohlin Fellowship at Emory Law School.

---

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

- |   |   |
|---|---|
| 1—Deaths  | 6—New Appointments                                  |
| 2—Retirements                                   | 7—Leaves for Special Appointments (NOT Sabbaticals) |
| 3—Foreign Scholars (visiting the USA or Canada) | 8—Resignations                                      |
| 4—Promotions                                    | 9—Miscellaneous                                     |
| 5—Administrative Appointments                   |   |

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment; her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Production Editor, *American Economic Review*, Room 8279, Bunche Hall, University of California, Los Angeles, CA 90024.

# The American Economic Review

## PAPERS AND PROCEEDINGS

OF THE

Ninety-Fifth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

New York, New York, December 29-30, 1982



MAY 1983

# THE AMERICAN ECONOMIC ASSOCIATION

●Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Dues for 1983, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:

\$32.00 for regular members with rank of assistant professor or lower, or with annual income of \$15,350, or less;

\$38.40 for regular members with rank of associate professor, or with annual income of \$15,350 to \$25,600;

\$44.80 for regular members with rank of full professor, or with annual income above \$25,600;

\$16.00 for junior members (registered students). Certification must be submitted yearly.

Subscriptions (libraries, institutions, or firms) are \$100.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$9.20 to cover extra postage.

●Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212.

Founded in 1885

## Officers

### *President*

W. ARTHUR LEWIS  
Princeton University

### *President-Elect*

CHARLES L. SCHULTZE  
The Brookings Institution

### *Vice Presidents*

JUANITA M. KREPS  
Duke University  
EDMUND S. PHELPS  
Columbia University

### *Secretary*

C. ELTON HINSHAW  
Vanderbilt University

### *Treasurer*

RENDIGS FELS  
Vanderbilt University

### *Managing Editor of The American Economic Review*

ROBERT W. CLOWER  
University of California-Los Angeles

### *Managing Editor of The Journal of Economic Literature*

MOSES ABRAMOVITZ  
Stanford University

## Executive Committee

### *Elected Members of the Executive Committee*

ELIZABETH E. BAILEY  
Civil Aeronautics Board  
ROBERT J. GORDON  
Northwestern University  
ANN F. FRIEDLAENDER  
Massachusetts Institute of Technology  
JOSEPH E. STIGLITZ  
Princeton University  
WILLIAM D. NORDHAUS  
Yale University  
A. MICHAEL SPENCE  
Harvard University

### *EX OFFICIO Members*

WILLIAM J. BAUMOL  
Princeton University and New York University  
GARDNER ACKLEY  
The University of Michigan

# THE AMERICAN ECONOMIC REVIEW

---

VOL. 73 NO. 2

MAY 1983

---

## *PAPERS AND PROCEEDINGS*

OF THE

*Ninety-Fifth Annual Meeting*

OF THE

AMERICAN ECONOMIC ASSOCIATION

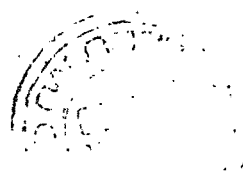
New York, New York

December 28–30, 1982

*Program Arranged by W. Arthur Lewis*

*Papers and Proceedings Edited by John G. Riley and Wilma St. John*

Copyright © American Economic Association, 1983



## CONTENTS

Editors' Introduction .....	<i>John G. Riley and Wilma St. John</i>	vii
-----------------------------	---	-----

## PAPERS

<b>Richard T. Ely Lecture</b>		
Monetary Policy and Economic Activity: Benefits and Costs of Monetarism .....	<i>Andrew F. Brimmer</i>	1
<b>Research in Economic Education</b>		
The Efficacy of Innovative Teaching Techniques in Economics: The U.K. Experience .....	<i>Keith G. Lumsden and Alex Scott</i>	13
Modeling Multiple Outputs in Educational Functions ...	<i>John F. Chizmar and Thomas A. Zak</i>	18
Who Maximizes What? A Study in Student Time Allocation .....	<i>Robert M. Schmidt</i>	23
<b>Economics of Fertility</b>		
Mortality Rates, Mortality Events, and the Number of Births .....	<i>Randall J. Olsen</i>	29
Economic Analyses of the Spacing of Births .....	<i>John L. Newman</i>	33
Consumer Demand and Household Production: The Relationship Between Fertility and Child Mortality .....	<i>Mark R. Rosenzweig and T. Paul Schultz</i>	38
<b>International Dimensions of Monetary Management</b>		
U.S. Monetary Policy and World Liquidity .....	<i>Thomas D. Willett</i>	43
Monetary Policy: Domestic Targets and International Constraints .....	<i>Jacob A. Frenkel</i>	48
Internationally Managed Moneys .....	<i>George M. von Furstenberg</i>	54
<b>Exploring Black Welfare Dependency</b>		
Changes in Black Family Structure: Implications for Welfare Dependency .....	<i>William Darity, Jr. and Samuel L. Myers, Jr.</i>	59
Budget Cuts as Welfare Reform .....	<i>Sheldon Danziger</i>	65
<b>Investment, Savings, and Incentives</b>		
The Determinants of Investment: Another Look .....	<i>Ben S. Bernanke</i>	71
Welfare Aspects of Current U.S. Corporate Taxation .....	<i>Alan J. Auerbach</i>	76
National Savings and Economic Policy: The Efficacy of Investment vs. Savings Incentives .....	<i>Laurence J. Kotlikoff</i>	82
<b>Recent Structural Change in the Capital Markets</b>		
The Process of Financial Innovation .....	<i>William L. Silber</i>	89
Policy Implications of Structural Changes in Financial Markets .....	<i>Edward J. Kane</i>	96
Financial Innovation in Canada: Causes and Consequences .....	<i>C. Freedman</i>	101
<b>The Role of Alien Entrepreneurs in Economic Development</b>		
An Entrepreneurial Problem .....	<i>Peter Kilby</i>	107
Chinese Entrepreneurs in Southeast Asia .....	<i>Yuan-li Wu</i>	112
The Levantines in Latin America .....	<i>William Glade</i>	118

**Women and Health**

Women and Absenteeism: Health or Economics? . . . . .	<i>Lynn Paringer</i>	123
Women and the Use of Health Services . . . . .	<i>Gail R. Wilensky and Gail Lee Cafferata</i>	128
Time Allocation, Market Work, and Changes in Female Health . . . . .	<i>Barbara Wolfe and Robert Haveman</i>	134

**Long Waves in Economic Activity**

Long Waves and Technological Innovation . . . . .	<i>Edwin Mansfield</i>	141
Long Waves and Economic Growth: A Critical Appraisal . . . . .	<i>Nathan Rosenberg and Claudio R. Frischtak</i>	146
Long Swings and the Nonreproductive Cycle . . . . .	<i>David M. Gordon, Thomas E. Weisskopf, and Samuel Bowles</i>	152

**The Changing Fortunes of Regions**

On the Effects of Federal Aid . . . . .	<i>George Tolley, Ronald Krumm, and Jeffrey Sanders</i>	159
Industrial Bases and City Sizes . . . . .	<i>J. Vernon Henderson</i>	164
Economists, Economics, and State Economic Policy . . . . .	<i>Roger J. Vaughan</i>	169

**Economics of Mass Migration from Poor to Rich Countries**

Leading Issues of Fact and Theory . . . . .	<i>Michael J. Greenwood</i>	173
International Migration Models and Policies . . . . .	<i>Edwin P. Reubens</i>	178
Trade Theory, Distribution of Income, and Immigration . . . . .	<i>Francisco L. Rivera-Batiz</i>	183

**Deregulation, Competition, and Efficiency**

Marginal vs. Average Cost Pricing in the Presence of a Public Monopoly . . . . .	<i>Donald J. Brown and Geoffrey M. Heal</i>	189
The Present Direction of the FCC: An Appraisal . . . . .	<i>Nina W. Cornell and Douglas W. Webbink</i>	194
The First Step in Bank Deregulation: What about the FDIC? . . . . .	<i>John H. Kareken</i>	198

**R&D and Productivity Increases**

Technological Change and Market Structure: An Empirical Study . . . . .	<i>Edwin Mansfield</i>	205
R&D and Productivity Growth: Policy Studies and Issues . . . . .	<i>Rolf Piekarz</i>	210
R&D and Declining Productivity Growth . . . . .	<i>F. M. Scherer</i>	215

**Macroeconomics: Major Issues and Developments**

Is Unemployment a Macroeconomic Problem? . . . . .	<i>Robert E. Hall</i>	219
Microeconomic Developments and Macroeconomics . . . . .	<i>Milton Harris and Bengt Holmstrom</i>	223
Is There a Monetary Business Cycle? . . . . .	<i>Christopher A. Sims</i>	228

**The World Food Situation**

Changing Trends in World Food Production and Trade . . . . .	<i>G. Edward Schuh</i>	235
Food Prospects for the Developing Countries . . . . .	<i>John W. Mellor</i>	239
Food Prospects in the Developing Countries: A Qualified Optimistic View . . . . .	<i>Malcolm D. Bale and Ronald C. Duncan</i>	244

**Segmented Labor Markets**

Labor Market Segmentation: To What Paradigm Does It Belong? . . . . .	<i>Michael J. Piore</i>	249
Segmented Labor Markets in LDCs . . . . .	<i>Dipak Mazumdar</i>	254
The Internalization of Labor Markets: Causes and Consequences . . . . .	<i>Bernard Elbaum</i>	260

**Recent Advances in the Theory of Industrial Structure**

Raising Rivals' Costs . . . . .	<i>Steven C. Salop and David T. Scheffman</i>	267
The Welfare Effects of Intermittent Interruptions of Trade . . . . .	<i>Glenn C. Loury</i>	272
Information, Competition, and Markets . . . . .	<i>Barry J. Nalebuff and Joseph E. Stiglitz</i>	278

**The Budget and Inflation**

Budget Expansion and Cost Inflation .....	<i>Assar Lindbeck</i>	285
The Interconnection Between Public Expenditure and Inflation in Britain .....	<i>Walter Eltis</i>	291
Money, Credit Constraints, and Economic Activity .....	<i>Alan S. Blinder and Joseph E. Stiglitz</i>	297

**Classical Economics: The Subsistence Wage, and Demand-Supply Analysis**

Marx and the Iron Law of Wages .....	<i>William J. Baumol</i>	303
The Classical Theory of Wages and the Role of Demand Schedules in the Determination of Relative Prices .....	<i>Pierangelo Garegnani</i>	309
On the Interpretation of Ricardian Economics: The Assumptions Regarding Wages .....	<i>Samuel Hollander</i>	314

**Chinese Economic Reforms**

Price Adjustment, the Responsibility System, and Agricultural Productivity ..	<i>Thomas B. Wiens</i>	319
Economic Reforms and External Imbalance in China, 1978-81 .....	<i>Bruce L. Reynolds</i>	325
Enterprise-Level Reforms in Chinese State-Owned Industry .....	<i>William Byrd</i>	329

**International Trade**

De-Skilling, Skilled Commodities, and the NICs' Emerging Competitive Advantage .....	<i>Alice H. Amsden</i>	333
Linking Up to Distant Markets: South to North Exports of Manufactured Consumer Goods .....	<i>Donald B. Keesing</i>	338
New Theories of Trade Among Industrial Countries .....	<i>Paul Krugman</i>	343

**The IMF and Conditionality**

Lender of Early Resort: The IMF and the Poorest .....	<i>G. K. Helleiner</i>	349
On Seeking to Improve IMF Conditionality .....	<i>John Williamson</i>	354
Devaluation: A Critical Appraisal of the IMF's Policy Prescriptions .....	<i>Louka T. Katseli</i>	359

**Special Report on Book Publication**

On Contracting with Publishers: Author's Information Updated .....	<i>Martin Shubik, Peggy Heim, and William J. Baumol</i>	365
--	---	-----

## PROCEEDINGS

Minutes of the Annual Meeting .....	385
-------------------------------------	-----

Minutes of the Executive Committee Meetings .....	387
---	-----

## Reports

Secretary .....	<i>C. Elton Hinshaw</i>	393
Treasurer .....	<i>Rendigs Fels</i>	397
Finance Committee .....	<i>Rendigs Fels</i>	398
Managing Editor, <i>American Economic Review</i> .....	<i>Robert W. Clower</i>	401
Managing Editor, <i>Journal of Economic Literature</i> .....	<i>Moses Abramovitz</i>	407
Director, <i>Job Openings for Economists</i> .....	<i>C. Elton Hinshaw</i>	409
International Economic Association .....		411
Committee on Economic Education .....	<i>Allen C. Kelley</i>	412
Representative to the National Bureau of Economic Research .....	<i>Carl F. Christ</i>	413
Representative to UNESCO .....	<i>Walter S. Salant</i>	415
Committee on U.S.-China Exchanges .....	<i>Gregory C. Chow</i>	417
Economic Institute Policy and Advisory Board .....	<i>Edwin S. Mills</i>	418
Committee on the Status of Women in the Economics Profession .....	<i>Elizabeth E. Bailey</i>	419

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

## Editors' Introduction

This volume contains the *Papers and Proceedings* of the ninety-fifth annual meeting of the American Economic Association.

The *Proceedings* record the business activities of the Association in 1982: the annual membership meeting; the March and December meetings of the Executive Committee; reports of the Association's officers and committees.

The *Papers* constitute the greater part of the volume. They comprise sixty-seven contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. The procedure governing selection of contributions for the *Papers* differs radically, of course, from that governing selection for the *American Economic Review*. About a year in advance, the Association's President-elect (in 1982, W. Arthur Lewis; in 1983, Charles L. Schultze), acting as program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* is published annually in the December issue of the *AER*.) The program chairman sets limits on the length of papers at various sessions, and invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions are to be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing twenty-three sessions, although a total of ninety-eight sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies. There is no standard practice with regard to the publication of comments and discussions, and each program chairman must de-

cide how to allocate available publication space between invited papers and discussions. In the present volume, as in last year's volume, however, we are publishing nothing but invited papers.

The rules under which papers are published in the *Papers and Proceedings* are also different from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be less than 3,000 words in three-paper sessions and less than 4,000 words in two-paper sessions. Second, papers are not subjected to any refereeing process. Third, their content and range of subject matter reflect the wishes of the program chairman to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

While authors are encouraged to submit their manuscripts earlier, in practice most are submitted at the meeting itself. Thus, there is no time for communication with every author about editing changes made in order to improve content and style and to satisfy space restrictions.

Rather than reject a paper because it is too long, every effort is made to reduce its length and, at the same time, preserve the main ideas. However, if such cuts do not seem feasible, we may ask the author to allow its consideration for publication in a regular issue of the *Review* subject to the usual refereeing process, or the author may be asked to withdraw the paper and submit it elsewhere. A paper is also rejected if, after reading it, we conclude that it is utterly without merit. This year we are pleased to report that no paper has been rejected on either ground.

JOHN G. RILEY  
WILMA ST. JOHN

## RICHARD T. ELY LECTURE

# Monetary Policy and Economic Activity: Benefits and Costs of Monetarism

By ANDREW F. BRIMMER\*

On October 6, 1979, the Federal Open Market Committee (FOMC) adopted an essentially monetarist operating approach to monetary policy. The key element was the decision to put more emphasis on influencing bank reserves and the monetary aggregates (which were already being targeted) and less on interest rates.

Three years later, on October 5, 1982, the monetarist approach was laid aside—at least temporarily. During that three-year experiment with monetarism, the Federal Reserve's policy achieved a substantial benefit, but it also imposed considerable costs: it helped to check inflation while causing a long period of economic stagnation and a sharp rise in unemployment.

In this lecture, the evolution of monetarism in the Federal Reserve is traced briefly. It is shown that the resort to closer monetary targeting reflected a pragmatic move by the Federal Reserve to moderate actual inflation and to erase deeply embedded inflationary expectations.

The impact of monetarism on the economy as a whole and on several of its major sectors is evaluated. The effects are visible in the financial markets (particularly on nominal and real interest rates) and on the real economy (especially on housing, capital formation, and unemployment).

The benefits of moderating inflation (and the dampening of inflation expectations) are shown. Finally, since the Federal Reserve may well return to a monetarist approach in the coming year, an effort is made to assess the economic consequences of such a move.

### I. Monetarism in the Federal Reserve: Triumph and Retreat

The adoption of a monetarist approach by the FOMC was the culmination of a drift that had been underway for nearly two decades.<sup>1</sup> However, in January, 1970, because of the shortfall between the FOMC's policy objectives and actual performance, an overt move was made to encourage the growth of the monetary aggregates. And in March of that year, the FOMC voted explicitly to make the money supply and bank credit the primary targets of open market operations.<sup>2</sup>

The next phase in the evolution of monetarist practice in the Federal Reserve was imposed by Congress. In March, 1975, in H. Cong. Res. 133, the House and Senate requested the Federal Reserve Board to report to Congress the targets set for the annual growth rates of the monetary aggregates. Subsequently, this requirement was mandated by the Full Employment and Balanced Growth Act of 1978 (also known as the Humphrey-Hawkins Act), which was a formal amendment to the Federal Reserve Act.

#### A. Analytical Basis of Monetary Targeting

In deciding to pursue monetary targeting, the Federal Reserve finally accepted the view that "money does matter" with respect to the performance of the national economy. Yet, the Committee and its staff (along with monetary economists generally) have con-

\*President, Brimmer & Company, Inc. From March, 1966 through August, 1974, I was a member of the Board of Governors of the Federal Reserve System.

<sup>1</sup>For a discussion of these trends, see my 1972 article.

<sup>2</sup>The quest for a monetary target is described by former Federal Reserve Governor Sherman J. Maisel in his 1973 book.

tinued to be troubled by uncertainty over the precise way in which money affects economic activity.

But after weighing a number of alternatives, the FOMC settled on a narrow definition of the money supply ( $M1$ , consisting of currency and transactions deposits) as the control variable. It was assumed that the amount of currency and transactions deposits required to finance a given amount of national income can be predicted with considerable precision. Consequently, if the growth of the narrow money supply is restrained, the expansion of nominal income over a specified period of time will likewise be constrained. This, in turn, will bring the rise in nominal income in line with the increase in the country's capacity to produce. The end result is to promote price stability.

Yet, over the last decade (and especially since the mid-1970's), the linkage between the narrowly defined money supply and income—given the level and configuration of interest rates—has lost some of its cohesion.<sup>3</sup> This weakening in the relationship appears to be the result of reactions of households and firms to the financial innovations induced by the high level of interest rates (judged against historical experience) that has prevailed in recent years. The latter represented a high opportunity cost of holding sterile deposits, so the public shifted excess balances into earning assets—particularly into NOW accounts, money market funds, and short-term marketable securities. After a considerable time lag, the regulatory agencies (and eventually Congress itself in the Monetary Control Act of 1981) took steps to facilitate the adjustment to these changes in the public's management of its liquid assets.

From the vantage point of monetary policy, these financial innovations represented an erosion of the underpinnings on which targeting of the narrow money supply was based. In fact, from the mid-1970's, the accumulating evidence suggested that the public's demand for narrowly defined money had shifted downward. This means that, for

a given level of nominal income and interest rates, the public was prepared to hold less money than earlier established historical relationships would have indicated. This meant that the demand for money is far less stable than the advocates of monetarism have asserted.<sup>4</sup>

More recent evidence has reinforced this conclusion. It now seems clear that the strong preference for currency and demand deposits recorded since the early months of 1982 reflected an increase in the precautionary demand for money—part of which the Federal Reserve chose to accommodate. Failure to have done so would have resulted in even higher interest rates.<sup>5</sup>

#### B. Adoption of Pragmatic Monetarism, 1979

The Federal Reserve's adoption of a monetarist approach in October 1979 was prompted by its failure to achieve the goals set in the previous year. These were to foster moderate economic expansion at home; check the rate of domestic inflation, and strengthen the international position of the dollar. To implement the policy, the range for the growth of  $M1$  (from the fourth quarter of 1978 to the fourth quarter of 1979) was set at  $1\frac{1}{2}$  to  $4\frac{1}{2}$  percent. For  $M2$ , the range was 5 to 8 percent.

However, because of a combination of factors, the Federal Reserve failed to achieve these objectives. For example, in the twelve months ending on September 26, 1979 (just prior to the adoption of the new technique),  $M1$  expanded by 4.7 percent. But in the thirteen weeks ending on the reference date, the seasonally adjusted annual rate (SAAR)

<sup>4</sup>This issue and suggested explanations of the problem are examined by Thomas Simpson and Richard Porter in their 1980 article.

<sup>5</sup>For an analysis of the data supporting this point, see Congressional Budget Office, *The Economic and Budget Outlook: An Update*. These calculations (based on statistical money-demand equations) indicate that short-term interest rates may have been between 2.4 and 3.4 percentage points lower over the second quarter of 1982 if the Federal Reserve had accommodated fully the apparent shift in the demand for liquidity. (See fn., p. 49.)

<sup>3</sup>For a discussion of these problems and efforts within the Federal Reserve to resolve them, see Stephen Axilrod's 1982 article.

was 9.0 percent. The more broadly based *M2* also recorded a growth rate substantially in excess of the FOMC's targets.

At the same time, the high rates of inflation continued (for example, the Consumer Price Index rose by 12.9 percent—*SAAR*—in the third quarter). This helped to strengthen expectations of future inflation—reflected in expanded speculative activity in domestic markets (including home buying), and the continued erosion of the international position of the dollar. These developments combined to force the Federal Reserve to adopt more drastic measures on October 6.

The key elements were: the discount rate was raised from 11 percent; an 8 percent marginal reserve requirement was set on increased "managed liabilities" of member banks; and the operating focus of monetary policy was changed to put more emphasis on bank reserves (and therefore the money supply) and less on day-to-day variations in short-term interest rates.

Among these measures, the third was the most crucial for the long run. It represented the adoption of a practical form of monetarism as a guide in the conduct of monetary policy.<sup>6</sup> At the same time, the Federal Reserve recognized that the emphasis on controlling the growth of bank reserves would lead to wider fluctuations in short-term interest rates.

### C. Credit Controls: Digression from Monetarism

However, despite the prompt implementation of the new monetarist approach, an increasing segment of the public became convinced that the rate of inflation would accelerate. This strengthening of inflationary expectations led to intense speculation in commodities, especially in precious metals.

<sup>6</sup>In passing, it might be noted that the decision was not based on a full-fledged inquiry of the Maisel variety that served as the basis for the initial introduction of monetary aggregates in the FOMC's deliberations in 1970. Instead, the Committee apparently drew on a summary of the conclusions derived from the existing body of analyses undertaken by the Federal Reserve staffs over the intervening years.

While the Federal Reserve was pursuing a restrictive monetary policy (and urging a more restrictive fiscal policy), there was also rising support for mandatory wage and price controls. However, both the Carter Administration and the Federal Reserve stood firmly against use of such controls. Yet, the Administration, succumbing to the political pressure "to do something" about inflation, adopted direct credit controls in mid-March, 1980, and President Carter delegated their implementation to the Federal Reserve Board. The specific controls were both complex and comprehensive. But the essential elements were higher marginal reserve requirements for commercial banks, consumer credit lenders, and money market funds.<sup>7</sup>

This resort to credit controls was clearly a digression from monetarism. But it also reflected a failure to appreciate the effects on the economy of the considerable monetary restraint the Federal Reserve had already applied. In fact, the mainsprings of economic growth were already weakening, and the shock of credit controls hastened the coming of recession and made the contraction the sharpest since 1937.

As the Federal Reserve began lifting the controls in the summer of 1980, the reactions of market participants outran the System's intent, and interest rates dropped further. At the same time, the Federal Reserve did delay the resumption of monetary restraint until the fall, when the pressure exerted against the growth of bank reserves became particularly severe.

### D. Conflicting Goals: Inflation vs. Recession

One result of the severe money and credit restraint was another drop in real output in

<sup>7</sup>In passing, I must observe that this use of marginal reserve requirements to influence the differential cost and availability of credit was quite similar to a proposal I made in 1970 when I was a member of the Federal Reserve Board. Several bills were introduced in Congress to implement the proposal, but the Board declined to support any of them on votes ranging from 4 to 3 to 6 to 1. An amplified exposition is also presented in my earlier paper (1975).

the first quarter of 1981. But, unlike its behavior during recessions in the past, the Federal Reserve did not act explicitly to change its basic policy of monetary restraint. In fact, in early October 1981, when the unfolding evidence presented by the FOMC staff pointed clearly to an impending (and possibly quite serious) recession, the Committee consciously chose not to seek faster growth of the money supply (measured by *M1-B*).<sup>8</sup> This decision was taken despite the fact that the growth of the money supply was running substantially below target (for example, 1.0 percent in the first nine months of the year compared with a target range of 3½ to 6 percent for the year as a whole).

### E. Retreat from Monetarism

The monetarist approach to the conduct of monetary policy began to be hedged in the summer of 1982. The initial tactical withdrawal was prompted by a strong public demand for liquidity which boosted the growth rates of all of the monetary aggregates above the targets set by the FOMC. In an appearance before the Senate Banking Committee in July, Federal Reserve Chairman Paul A. Volcker reiterated the FOMC's annual growth targets for 1982, but he then took a significant step back from strict adherence to monetarism. He stated that:

...[G]rowth somewhat above the targeted ranges would be tolerated for a time in circumstances in which it appeared that precautionary or liquidity motivations, during a period of economic uncertainty and turbulence, were leading to stronger-than-anticipated demands for money. We will look to a variety of factors in reaching that judgment, including such technical factors as the behavior of different components in the money supply, the growth of credit, the behavior of banking and financial markets, and more broadly, the behavior of velocity and interest rates. [1982b, p. 491]

<sup>8</sup>See policy record of the Federal Open Market Committee, October 5-6, 1981 as published in the *Federal Reserve Bulletin*, December 1981.

The retreat from monetarism was signalled formally at the FOMC meeting on October 5, 1982, and the key elements were made public by Chairman Volcker on October 9. He began with an explanation of the reduction of the discount rate to 9½ percent made on the previous day. The "...change was... designed to maintain an appropriate alignment with short-term market rates...[But] it was also...taken against a background of continued sluggishness in business activity, the exceptional recent strength of the dollar on the exchange markets, and indications of strong demands for liquidity in some markets" (1982a, p. 691).

Chairman Volcker then proceeded to describe a significant change in the strategy (but not fundamental objectives) of monetary policy:

...In assessing economic and financial developments over recent months... there is growing evidence that the inflationary momentum has been broken. Indeed, with appropriate policies, the prospects appear good for continuing moderation of inflation in the months and years ahead...

We express policy in terms of broad targets for the various definitions of money on the basic thesis that, over time, the inflationary process is related to excessive growth in money and credit. But [I] repeatedly express caution about the validity of any single measure, or even all the measures in the short-run... [1982a, p. 691]

This general doubt about the validity of the monetary aggregates (especially of *M1*) turned into certainty against the backdrop of both recent and prospective changes in the ways in which consumers could hold their liquid balances. Among such changes were the near term maturity of some \$31 billion of "All Savers Certificates" and the planned introduction of "money market fund type" deposit accounts by banks and thrift institutions. Because of these distortions, the money aggregates—at least for the time being—could not be depended on as reliable guides for monetary policy.

Finally, Chairman Volcker stated that, although the various monetary aggregates already had been growing at rates somewhat above the paths targeted for the year, the FOMC would not restrict the availability of bank reserves to force them back into their respective ranges. The basic reason was clear: an accommodative monetary policy would be pursued in order to help check the deepening recession and to promote economic recovery.

So, at least for the near-term, the hands of the monetarists would no longer guide the nation's monetary policy.

### *F. Laxity in Fiscal Policy*

While the Federal Reserve was attempting to use monetary restraint to check inflation, it was getting virtually no help from the federal government's fiscal policy. For example, between 1961 and 1982, the federal government ran a budget deficit in every year except 1969. Moreover, partly because of the prolonged recession but also because of substantial legislated reductions in tax revenue, the federal budget is likely to carry a structural deficit beyond the middle of this decade.

In fiscal year 1979, federal outlays amounted to \$491.0 billion, and receipts were \$463.3 billion—leaving a deficit of \$27.7 billion. Outlays and the deficit represented 20.3 and 1.1 percent of *GNP*, respectively. Over the next two fiscal years (partially because of two recessions back-to-back), federal budget outlays rose to \$657.2 billion in 1981, and the deficit amounted to \$57.9 billion. In that year, the budget was 22.4 percent—and the deficit 2.0 percent—of *GNP*.

In 1982, expenditures rose by 10.8 percent to \$728.4 billion, but receipts increased only 3.1 percent to \$617.8 billion. So the deficit jumped by 91.2 percent to \$110.7 billion. These changes raised outlays to 23.8 percent of *GNP* while the deficit climbed to 3.6 percent of total output. The budget adopted by Congress for fiscal 1983 projects budget outlays at \$769.8 billion and receipts at \$665.9 billion, leaving a deficit of \$103.9 billion. These estimates would leave budget outlays and the deficit at 23.4 and 3.2 per-

cent of *GNP*, respectively. On the other hand, the unfolding evidence suggests strongly that both outlays and the deficit in fiscal 1983 will greatly exceed the amount anticipated in the budget legislation. Currently, estimates by private economists as well as by the Congressional Budget Office project fiscal 1983 outlays in the range of \$840–865 billion and the deficit in the range of \$175–200 billion. At these levels, federal outlays would represent between 25.6 and 26.3 percent of *GNP*, and the budget deficit would be equal to 5.3–6.1 percent of total output.

These large and persistent federal budget deficits added to aggregate demand and undoubtedly contributed to inflationary pressures in the economy at least through 1979. On the other hand, during the period of substantially reduced economic activity over the last two years, the deficit helped to prevent demand from falling as much as cutbacks in private spending would have brought about.

But, in the context of monetary policy, the most important aspect of large budget deficits is the pressure they exert on the money and capital markets. For example, in 1979, net borrowing by the U.S. Treasury amounted to \$31.0 billion—representing 7.8 percent of the total funds raised in that year. In 1980, the net amount jumped to \$75.6 billion and absorbed 21.6 percent of the total. In that year, the amount raised by the private sector (including state and local governments) dropped by \$92.5 billion as the recession cut the demand for funds. In 1981, net borrowing by the federal government remained essentially unchanged at \$77.8 billion. Total funds raised increased to \$408.7 billion, so the federal government's share eased slightly to 19.0 percent.

Federal government deficit financing increased dramatically in 1982. Net borrowing by the U.S. Treasury is estimated at \$148.2 billion while the amount of funds raised by all sectors is projected at \$399.4 billion. If these estimates materialize, the federal government would have absorbed 37.1 percent of the total funds raised in the capital market. Moreover, net government borrowing in 1983 is projected at \$186.2 billion, and total borrowing is estimated at \$456.2 bil-

lion. Under this scenario, the federal government would still be absorbing 36.9 percent of the total funds raised by all economic sectors.

This very large anticipated demand for funds by the federal government has cast a shadow over the nation's financial markets. Contemplating it, investors in government securities (and this is especially true of holders of long-term bonds) have modified their portfolio strategy accordingly. They expect the heavy flow of government bonds to exert upward pressure on long-term interest rates. This has induced them to concentrate a disproportionate share of their currently available funds in short- and medium-term issues. This action helps to validate their expectations of higher yields on long-term securities in the near-term as well as in the future.

## II. Empirical Assessment of Monetary Restraint: Analytical Techniques

To assess the principal impact of monetarism on the nation's economy, the effects of monetary targeting were simulated with the Data Resources, Inc. (DRI) quarterly econometric model of the United States. Additional calculations were made by Brimmer and Company, Inc., using its own analytical tools.

In the first evaluation of monetarism, the simulation of the American economy with the DRI model made it possible to compare the pattern of economic activity in 1979 and 1980 under the influence of the new series of measures adopted in those years with what it would have been otherwise. The DRI control forecast of August 24, 1980 was used as baseline. The simulation was designed to replicate history without the shift to the monetarist approach.<sup>9</sup>

<sup>9</sup>In technical statistical terms, the historical patterns of economic behavior were traced by feedback of residual errors prior to the third quarter of 1980. In fact, the actual pattern of economic activity from the fourth quarter of 1979 to the third quarter of 1980 incorporated the actual and estimated effects of the new monetary policy.

An alternative DRI simulation which incorporated a recession was used as a benchmark for 1981. This exercise captured the prospects for the economy in the summer of 1981—corresponding roughly to the point in time when the FOMC chose to maintain the prevailing degree of monetary restraint although a recession was clearly visible on the horizon.<sup>10</sup> The accommodative monetary policy alternative in 1981 is represented by the DRI control forecast prepared at the end of July. Roughly one year later, in the summer of 1982, the FOMC faced a parallel situation and reached a similar conclusion.<sup>11</sup> So, for 1982, the monetarist effects are captured by a recession simulation of the DRI model. For that year, the alternative is a DRI model solution which incorporates appreciably lower rates of inflation and of interest rates than embodied in the control forecast.

The final assessment of the probable impact of monetarism on the economy was prepared on the assumption that the Federal Reserve might return to the targeting of growth rates for the monetary aggregates fairly early in 1983. This possibility is captured in an alternative simulation of the DRI model prepared in early December, 1982, which results in substantially higher interest rates and pushes the economy back into recession during the second half of 1984. These results are compared with the DRI control forecast also prepared in early December, 1982.

## III. Impact of Monetarism: Financial Sector

The differential effects of monetarist policies on the financial sector can be traced in the behavior of bank reserves, monetary aggregates, and interest rates. The impact is felt first in the money market—and especially by commercial banks. After a time lag (which is not very long), the effects show up in the

<sup>10</sup>See Federal Open Market Committee policy record, 1981.

<sup>11</sup>See the policy record of the Federal Open Market Committee, June 30–July 1, 1982 and July 5, 1982 meetings as reprinted in *Federal Reserve Bulletin*, September 1982.

capital market where the behavior of long-term investors (such as life insurance companies and pension funds) is of major importance.

#### A. *Impact on the Money Market*

As indicated above, when the FOMC adopted its monetarist approach in October, 1979, it wanted to achieve a progressive reduction in the growth rate of both bank reserves and the money supply. The figures show the FOMC achieved both objectives.

For example, in the fourth quarter of 1979, adjusted bank reserves<sup>12</sup> grew at an annual rate of 7.7 percent, but by the second quarter of 1980, the rate had dropped to zero. In the aftermath of the 1980 recession, the growth rate spurted to 10.1 percent in the final quarter of that year. However, renewed restraint cut the growth rate to zero in the third quarter of 1981, and reserves actually shrank at an annual rate of 3.4 percent in the final quarter of that year. After a bulge (at an annual rate of 15.4 percent) in the first three months of 1982, the growth rate decreased progressively to 1.6 percent in the third quarter.

The behavior of the money supply essentially paralleled that of bank reserves. For instance, *M1* rose at an annual rate of 10.3 percent in the second quarter of 1979, and by 9.6 percent in the third quarter. Both figures were much above the FOMC's target. But with the sharp turn to restraint, the growth rate receded to 4.6 percent in the final three months of that year, and in the second quarter of 1980, the money supply shrank at an annual rate of 3.1 percent. Again, following the 1980 recession, the *M1* growth rate jumped to 14.8 percent in the third quarter. But in this case, also, the growth dropped off to 0.3 percent in the summer of 1981. After climbing to 10.9 percent in the first three months of 1982, the growth rate of *M1* receded to 3½ percent in both the second and third quarters.

<sup>12</sup> Consist of adjusted monetary base, less currency held by the nonbank public as calculated by the Federal Reserve Bank of St. Louis.

In 1979, the growth rates required by both accommodative and monetarist policies were essentially the same—differing by only 0.1 percent. However, as the monetarist policies took hold, the gap between the two widened—from 0.4 percent in 1980 to 0.9 percent in 1982.

#### B. *Effects on Interest Rates*

The immediate effect of the new policy on the money market was to increase the federal funds rate by 211 basis points over the rate that would have prevailed under the premonetarist regime. Over time, the spread narrowed, but it remained close to 1.0 percentage point during most of the last three years. Finally, as a by-product of the monetarist approach, the federal funds rate has been far more volatile than it was in the premonetarist era.

The effect of the new policies on commercial banks' prime lending rates was slightly less than that registered on federal funds. Nevertheless, the margin attributable to the new technique averaged about 50 basis points above accommodative rates in the 1979–82 period. Since 1979, yields on corporate bonds have averaged about 50 basis points above what an accommodative monetary policy would have produced. Finally, yields on both corporate bonds and U.S. government securities have become more volatile in the last three years than they were in 1976 through 1979.

### IV. *Economic Costs of Monetarism*

The monetarist policies followed by the Federal Reserve since 1979 have helped to keep the American economy operating substantially below its potential. In fact, the two policy-induced recessions—to which monetary policy made a significant contribution—have so restrained the growth of output and raised the level of unemployment that the economy can be accurately described as suffering from a mini-depression.

Measured in terms of real *GNP*, there has been no net growth in the United States in the last three years. In 1982, real output ran

about 0.2 percent below the level recorded in the first quarter of 1980. Moreover, when actual output is compared with the economy's capacity to produce at full employment, the depressed condition of the economy becomes even more apparent. In 1979, actual output was equal to 99.7 percent of potential, but by 1982 the proportion had dropped to 87.5 percent. This shortfall represents a loss of approximately \$450 billion in real output since 1979. In human terms, between 1979 and 1982, real GNP dropped from \$6,707 to \$6,348 per capita—a decline of 5.4 percent.

The monetarist policies have had a particularly adverse effect on the nation's housing sector. The annual level of housing starts dropped from 1.7 million in 1979 to just over 1.0 million in 1982. Over the three years 1980–82, slightly more than 3.4 million units were started. A more accommodative monetary policy would have raised starts to 3.7 million. To satisfy the long-term demand for housing, about 2.0 million units should be started each year. Against that benchmark, the shortfall in home production since 1979 amounts to roughly 2.6 million units—a loss equal to over two-fifths of potential output.

The linkage between monetarist policies and the housing sector is readily understood. Sharp rises in interest rates cut the inflow of funds to savings and loan associations which still serve as principal home financing institutions. In response, mortgage interest rates also rose dramatically. The increase in the latter outstripped the advance in disposable personal income, so the number of would-be home buyers who could qualify for mortgages shrank significantly.

The impact of the monetarist policies on capital formation in the business sector has also been dramatic. For instance, in 1979, in a sample of large manufacturing firms with high bond ratings, the cost of debt averaged about 10 percent, and the cost of equity was 15 percent. By 1981 (while the debt-equity ratios remained essentially unchanged at 25 and 75 percent, respectively), the cost of debt had climbed to 15 percent, and the cost of equity was up to 18 percent. So for this group of firms, the weighted average cost of capital increased from 13.8 to 17.3 percent over the course of two years. Over the same

period, cuts in capital outlays ranged between 30 and 40 percent compared to original plans.

The large increase in interest rates had a noticeably dampening effect on the rate of capital formation in the country at large. For instance, over the three years ending in 1979, real outlays for business fixed investment expanded at an annual rate of 2.8 percent. But over the following three years, the level of real investment was essentially stagnant. In addition to high interest rates, the large backlog of excess capacity in industry has also dampened the incentive to invest in new facilities. At the peak of economic activity in the first quarter of 1980, the capacity utilization rate in manufacturing was 83.4 percent (having receded from 86.7 percent a year earlier). But by November, 1982, the rate had dropped to a record low of 67.8 percent.

In the labor market, the adverse effects of monetarist policies are also plainly evident. In November 1982, the unemployment rate was 10.8 percent of the civilian labor force. Behind this ratio were roughly 12 million people without jobs. For the year as a whole, unemployment might average 10.7 million, and the unemployment rate might average 9.7 percent. In 1979, the corresponding figures were 6.1 million and 5.8 percent.

If the Federal Reserve had followed a more accommodative monetary policy over the intervening years, the unemployment rate in 1982 may have been roughly 7.0 percent. This rate would have translated into an unemployment level of 7.7 million. The latter estimate is roughly 1.0 million higher than it would have been if the 1979 unemployment rate had been maintained. On the other hand, the estimate is 3.0 million below the actual level of unemployment that is likely to have prevailed in 1982.

#### V. Benefits of Monetarism: Moderation of Inflation

The principal benefit of the Federal Reserve's restrictive monetary policy is a substantial moderation in inflation. In broad terms (as measured by the GNP implicit price deflator), the rate of inflation climbed from 5.2 percent in 1976 (in the wake of the

1973–75 recession) to 9.4 percent in 1981. And during the first three quarters of 1982, inflation averaged  $4\frac{1}{2}$  percent. The inflation profile traced by the Consumer Price Index (CPI) is even more striking. From a low of 5.8 percent recorded in 1976, this measure of inflation rose by 13.5 percent in 1980. It then slowed to 10.4 percent in 1981. During 1982, the expected increase is in the neighborhood of  $4\frac{1}{2}$  percent.

### A. Inflationary Expectations

Beyond these quantitative measures, considerable success has also been achieved in dampening deep-seated inflationary expectations. Although reported inflation spurted in 1979 and 1980, the forces which led to a climb in core inflation were at work for most of the last decade. Yet, it took time for the public to adapt to the new inflationary environment. This adjustment took numerous forms, but a strong preference for physical assets (especially housing) was particularly evident. Moreover, the public's reactions to inflation contributed to the stimulation of expectations of further inflation—which were subsequently validated.

Economists and businessmen—as well as public officials—have found it extremely difficult to understand or predict inflation. For instance, it is clear in retrospect that the Carter Administration seriously misjudged the degree to which inflation had become embedded in the economy. Until late in their tenure, they seemed to have expected inflation to average in the 5–7 percent range—rising to about 8 percent in a period of strong economic expansion and falling to around 5 percent as the economy slowed down. This relatively optimistic projection clearly led to the adoption of policies which resulted in the provision of too much federal government stimulation of the economy.

Subsequently, the Reagan Administration also misjudged the nature and tenacity of inflation. They assumed that a highly restrictive monetary policy would check inflation in a predictable way. Moreover, guided by a not well-defined “supply-side” view of the economy, they thought inflation could be stopped without seriously adverse effects on

output and employment. This same view led them to seek—and achieve—a very large reduction in tax rates. The latter was expected to stimulate risk taking and productive effort, and bring about a sizable (and early) expansion in economic activity. Partly because of this error in judgment (and especially because of the large budget deficits resulting from the tax reduction), the public's expectation of future inflation has been reinforced.

Part of the inflation experienced over the last decade can be traced to a series of supply shocks which arose from a combination of shortages, producer cartels, and substantial declines in the value of the dollar in the foreign exchange markets. The most pronounced shocks of all were the successive waves of oil price increases posted by the Organization of Petroleum Exporting Countries (OPEC).

The inflation resulting from shocks was aggravated by the conditions associated with excess demand in the economy as a whole. Whenever there was insufficient slack in the economy, firms took advantage of sellers' markets to raise prices, and trade unions raised wage demands. Both groups could generally anticipate that—with tight markets for goods and services—the general price level might rise by, say, 3 percent more than before. Consequently, instead of raising prices and wages by 3 percent, they sought to lift them by 6 percent in an effort to get a real increase of 3 percent.

### B. Inflation and Real Interest Rates

The long bout of inflation has led to a marked change in the behavior of investors in the long-term capital market. Partly because of the strong inflationary pressures, the *nominal* interest rate (i.e., the interest rate stated as a percentage of the principal) rose sharply. But, what is important for both borrowers and lenders is the *real* interest rate—that is, the interest rate adjusted for inflation.

Over the last several years, real rates have risen substantially. If such rates are measured by the differential between yields on long-term U.S. government bonds and the

rate of inflation (measured by the *GNP* deflator), the real rate climbed from 1.18 percent in 1978 to 6.87 percent in 1982. If the *CPI* is used as a benchmark, real interest rates rose from .78 percent in 1978 to 6.7 percent in 1982. The increase in real rates was even larger in the case of yields on corporate bonds. In this case, the jump was from 1.58 percent in 1978 to 7.88 percent in 1982 (when measured on the basis of the *GNP* deflator). Using the *CPI*, the rise was from 1.8 to 7.61 percent over the same period.

The effects of real interest rates on investment decisions can be seen readily. For instance, if the nominal rate is 8 percent during a period in which prices are rising at a 7 percent rate, this 8 percent yield represents only a 1.0 percent real rate. This would also represent an unusually low cost of borrowing, but it would also represent an unusually low rate of return to investors. An investor who lends \$100 at 8 percent for one year is being paid back in real terms only \$93, which together with the \$8 interest receipts amounts to a real interest receipt of only \$1.

Furthermore, most investors do not focus on the real rate of interest before taxes, but on the real rate after any income tax has been paid. In the example cited above, if both the lender and the borrower are in the 50 percent tax bracket, the lender gets back after taxes a principal of \$93 plus \$4 (\$8 times 50 percent) of after-tax interest income—that is, a total of \$97. For this lender, the real after-tax interest rate is negative. By contrast, the borrower makes an extra after-tax payment of only \$4 and gains \$7 from the fall in the real value of the principal. So, for the borrower, the before-tax nominal interest rate of 8 percent is a real after-tax cost of borrowing of *minus* 3 percent.

The statistical data show that inflation does raise the nominal before-tax interest rate, and it may also increase the real rate after allowing for taxes. To illustrate, assume that initially prices are stable and the nominal interest rate is 3 percent. Assume also that prices now rise at a 7 percent annual rate and the marginal tax rate of borrowers and lenders is 50 percent. Under these conditions, a 17 percent nominal before-tax interest rate would be required to leave borrowers

and lenders with the same 1.5 percent real after-tax rate they had before the inflation.

Finally, under the impact of the prolonged inflation (and partly because the Federal Reserve began to concentrate more directly on the control of bank reserves and the money supply), interest rates have become far more volatile, and capital markets have become less stable. As a consequence, uncertainty has increased considerably, and investors are demanding a large premium in the form of higher rates of return if they are to undertake the increased risk associated with financial commitments—especially in the long-term capital market.

## VI. Future of Monetarism

In attempting to assess the future of monetarism in central banking in this country, it is well to remember the Federal Reserve's repeated statement that it has *not* changed monetary policy. It stresses that it will continue its campaign to check inflation and to eradicate the still deeply rooted inflationary expectations. Therefore, the only real questions concern the degree of restraint which the Federal Reserve will exert in the future and the nature of the tools it will employ in the pursuit of its goals.

That being the case, the possibility of a return to enforced monetary targeting in early 1983 remains quite alive. Yet, if the Federal Reserve were to return to a fundamentally monetarist approach to policy, the adverse effects on the economy would be substantial.

For example, the growth rate of the money supply (measured by *M2*) might be kept in the neighborhood of 5 to 6 percent in 1983 and 1984. These rates of expansion would be 2 to 3 percentage points below what would be required by an accommodative monetary policy.

This degree of monetary restraint would give a fairly sharp boost to interest rates. The federal funds rate might average between 10 and 11 percent in 1983 and between 12 and 13 percent in 1984. These levels would be at least 2 to 3 percentage points higher than those that might result from a continuation of the Federal Reserve's present tolerance for over-target growth of the monetary aggre-

gates. Under a restored monetarist regime, the commercial banks' prime rate might average between 12 and 13 percent in 1983 and between 14 and 15 percent in the following year. Yields on high-grade corporate bonds might be in the range of 12 to 13 percent in 1983, and they may climb to between 13 and 14 percent a year later. An accommodative monetary policy might achieve long-term interest rates 3 to 4 percentage points below the level that might result under a monetarist regime.

A continuation of the Federal Reserve's liberalized monetary policy might still leave real interest rates on U.S. government bonds in the range of 5 to 6 percent in 1983 and between 4 and 5 percent in 1984. On high-grade corporate bonds, real interest rates might continue to range between 5 and 6 percent in both years. An accommodative monetary policy might cut these real rates in half.

#### A. Prospects for Economic Activity

Under a restored monetarist approach to policy, real *GNP* might rise by less than 2 percent in 1983, and no growth at all could be anticipated in the following year. In fact, sometime in 1984, another fairly serious recession would become a real possibility. This continued stagnation in economic activity would widen further the gap between potential *GNP* and actual output. The loss in real *GNP* over the two years might amount to \$470 billion. This figure would be the equivalent of  $13\frac{1}{2}$  percent of potential output. Such an outcome would mean very little improvement in real per capita income over the next several years.

In a labor market context, the return to a monetarist regime would mean continuous wastage of a large fraction of the nation's human resources. The level of unemployment would remain on a mini-depression scale. An unemployment rate of about 10 percent might be expected in 1983, and it might still be around 9 percent in 1984. These fractions would translate into  $11\frac{1}{2}$  million unemployed persons in 1983 and to just over 10 million in 1984. Moreover, an extremely restrictive monetary policy would

probably keep the unemployment rate in excess of 8 percent beyond the middle of this decade.

#### B. Outlook for Inflation

Over the next year or two, the recent abatement in inflationary pressures is not likely to be erased. The large measure of excess capacity in industry (which is likely to remain for some time) means that output can be increased substantially without running into the type of shortages and bottlenecks which would generate upward pressure on product prices. Moreover, the lower level of economic activity, not only in this country but in the world at large, will continue to dampen the demand for oil. This means that there is little prospect of a sharp boost to inflation from this sector.

In addition, the underlying rate of inflation (reflected in the tendency for increases in compensation to exceed gains in productivity) will probably remain moderate for quite some time. The major concessions on wages and benefits which numerous strong trade unions have made (mainly under recession-induced pressures) are not likely to be withdrawn quickly. Consequently, the rate of increase in unit labor costs will also most likely remain quite moderate.

In conclusion, the greatly improved prospect for inflation—and the continued uncertain outlook for the real economy—provides a considerable margin of safety within which the Federal Reserve can conduct an accommodative monetary policy well into next year. It also means that such a policy is not likely to rekindle the kind of strong inflationary expectations which the Federal Reserve has sought to erase over the last several years.

#### REFERENCES

- Axilrod, Stephen H., "Monetary Policy, Money Supply, and the Federal Reserve's Operating Procedures," *Federal Reserve Bulletin*, January 1982, 68, 13-24.
- Brimmer, Andrew F., "Central Banking and Credit Allocation," Bureau of Business Research, University of Texas-Austin,

1975.

\_\_\_\_\_, "The Political Economy of Money: Evaluation and Impact of Monetarism in the Federal Reserve System," *American Economic Review Proceedings*, May 1972, 62, 344-52.

\_\_\_\_\_, "Supplemental Reserve Requirements on Member Bank Assets," Statement before the Subcommittee on Financial Institutions of the Committee on Banking, Housing and Urban Affairs, U.S. Senate, April 7, 1971; reprinted in *Federal Reserve Bulletin*, April 1971, 57, 307-19.

Maisel, Sherman J., *Managing the Dollar*, New York: W. W. Norton & Co., 1973.

Simpson, Thomas D., and Porter, Richard D., "Some Issues Involving the Definition and Interpretation of Monetary Aggregates," *Controlling the Monetary Aggregates III*, Boston: Federal Reserve Bank of Boston, 1980, 161-234.

Volcker, Paul A., (1982a) Excerpt from an

informal talk to the Business Council at Hot Springs, Virginia, October 9, 1982; reprinted in *Federal Reserve Bulletin*, November 1982, 68, 691-92.

\_\_\_\_\_, (1982b) Statement before the Committee on Banking, Housing, and Urban Affairs, U.S. Senate, July 20, 1982; reprinted in *Federal Reserve Bulletin*, August 1982, 68, 487-94.

Board of Governors of the Federal Reserve System, "Record of Policy Actions of the Federal Open Market Committee," *Federal Reserve Bulletin*, December 1981, 67, 905-10.

\_\_\_\_\_, "Record of Policy Actions of the Federal Open Market Committee," *Federal Reserve Bulletin*, September 1982, 68, 541-49.

Congressional Budget Office, "Evidence of Money Demand Instability in 1982," *The Economic and Budget Outlook: An Update*, Appendix A, September 1982, 79-85.

## RESEARCH IN ECONOMIC EDUCATION

# The Efficacy of Innovative Teaching Techniques in Economics: The U.K. Experience

By KEITH G. LUMSDEN AND ALEX SCOTT\*

During the academic year, 1979–80, over 2,500 students studying first-year economics in nineteen U.K. universities and polytechnics were involved in a research project, the aim of which was to attempt to assess the efficacy of innovative teaching techniques in basic economics. The new techniques included were TIPS (see Allen C. Kelley, 1968), Cases, Programmed Learning, and Macro-simulations.

Course packages were constructed, combining innovative and conventional techniques in different proportions. Each course package was designed within the overall objectives of a research design intended to generate a data matrix with sufficient observations in each cell to test the impact of various teaching techniques on different types of students in different institutional settings. Three conventional courses which did not utilize the innovative techniques were included to provide norming data.

The common three-hour final examination, which yielded several measures of output, consisted of 20 multiple choice questions to measure knowledge of concepts and simple to intermediate applications, 1 problem-case to measure complex applications and analysis, 1 micro essay and 1 macro essay to measure synthesis and evaluation. There was no choice in the selection of questions to be answered.

To ensure that the case and essay marks were consistent across institutions, each paper was regraded by one experienced university lecturer and a sample regraded for a third time to test his consistency.

In the three years prior to the experimental year, a total of 26 pilot courses were run in seven institutions to solve logistics problems. Based on the pilot studies, participating institutions adopted one of three broad strategies: (i) the conventional course was scrapped and innovative techniques were substituted for tutorials and essays—"complete substitution"; (ii) conventional inputs were reduced and innovative techniques were substituted—"partial substitution"; (iii) innovative techniques were added to existing conventional inputs—"add on."

Since each teaching technique has a price tag, the strategies produced widely varying course costs, average total cost per student, and marginal cost. An indication of the scope for cost variation can be gained from the fact that the typical conventional first-year course containing 300 students would require 3 lectures and 30 tutorial hours per week (typically there are 10 students per tutorial). In contrast to the use of graduate students in beginning economics tutorials in the United States, tutorials in the United Kingdom are usually shared among all faculty members. A tutorial hour counts as a full contact hour in calculating teaching loads. Thus the opportunity cost of the conventional tutorial system in the United Kingdom could be as high as 7 or 8 upper level courses. Some innovative courses were implemented which scrapped the labor-intensive weekly tutorial system and substituted TIPS, Cases, and one or two hours of class remedial tutorials which students could attend at their own discretion.

A comparative institutional cost model was used to derive average and marginal costs. The model included faculty inputs at average U.K. rates, and assumed a 10 hours per week teaching load, and 30 percent of time devoted to research; items such as course as-

\*The Esmee Fairbairn Research Centre, Heriot-Watt University, Edinburgh. Our research was funded by the Department of Education and Science in the U.K., and The Esmee Fairbairn Charitable Trust.

TABLE 1—COURSE COSTS BY TYPE OF COURSE

Type of Course	Range of Cost (£1980)	
	ATC per Student	Marginal Cost
Complete Substitution (6 courses)	15–30	7–14
Partial Substitution (3 courses)	19–26	16–19
Add On (7 courses)	32–53	29–35
Conventional (3 courses)	29–44	25–34

sistants, duplication and computer costs were also included. The differences in average total cost per student and marginal cost by type of course are summarised in Table 1.

How well does the maxim “you never get more than you pay for” hold for beginning economics students in the United Kingdom?

To relate course inputs to course outputs, two techniques were used. First, every institution was ranked by its mean performance on each of the 10 output measures (i.e., the various parts of the final examination taken in different combinations and student opinion of the course) and compared with the input cost data. Second, multiple regression analysis was carried out in which each of the output measures was the dependent variable and in which student and course input characteristics were the independent variables.

Using the first approach, the most striking feature is the lack of any obvious connection between course inputs and outputs. Courses with high resource inputs per student are as likely to come at the bottom of the rankings as at the top, both in terms of the performance and opinion measures of output. Furthermore, courses move about in the rankings according to the measure of output used indicating that different measures of output yield different answers to questions of relative cost effectiveness.

The regression results presented some severe problems. (i) The coefficients of the set of variables relating to the course characteristics adopted odd values. For example, using the microeconomic essay as dependent variable (mean 46) the coefficient of using a

TIPS was 54.5, using Cases was –27.9, and using a Programmed Text was 27.3. Clearly, these coefficients were too large to reflect learning related effects. (ii) A slight alteration in the number of institutions included in the sample affected both the magnitude and the sign of the coefficients. (iii) A slight alteration in the specification of the equation, such as excluding variables relating to the textbooks used, completely altered the coefficients.

The regression results were therefore both unrealistic and highly unstable. Closer investigation led us to conclude that (i) although this was a large scale study in terms of the number of student observations and institutions involved, the number of observations on different courses was relatively small in relation to the set of variables required to represent each course package. In addition, because of the different course designs, each set of course-related variables was unique, that is, there was no precise duplication of course design between institutions. There were considerable differences in the means of the output variables between institutions (the microeconomic essay, for example, varied from 38.4 to 52.4); these differences were not eliminated by the inclusion of student-related variables, hence the unique group of course input variables was picking up the large interinstitution differences in an unpredictable manner. Subsequent simulations using variations in the number of institutions and different specifications demonstrate that any number of different “findings” can be generated from this data set; and that (ii) the factors affecting learning within institutions depended on the individual case. Selecting the one new teaching technique on which the largest amount of data is available (TIPS), Table 2 details, by institution, the statistically significant coefficients relating to the number of TIPS surveys done by each student and various measures of output.

From the eleven institutions, in only two are the coefficients significant for all output measures; in three institutions, there are no significant coefficients for any measures of output. In six institutions, the coefficients are significant for essay marks. While the

TABLE 2—NUMBER OF TIPS SURVEYS DONE AND MICRO EXAMINATION PERFORMANCE  
STATISTICALLY SIGNIFICANT REGRESSION COEFFICIENTS

Institution	Examination			Total Mark
	Multiple Choice	Case	Essay	
01	.09		.99	1.53
03	.24			
04				
08	.34			
09				
10	.08	2.00	1.29	2.39
11			1.78	
13	.07		.99	
14				
16			2.49	
24		1.77	2.29	3.28

TABLE 3—STUDENT COURSE GOALS

Course	Number of Students	Percentage Responding				
		1	2	3	4	5
Micro	1746	3	8	29	31	29
Macro	1217	6	9	32	30	23

number of observations are much smaller for students using cases and macrosimulations, similar findings emerge.

These two statistical problems—the instability of the course input variable coefficients and the unique nature of learning effects within institutions—must raise questions about the validity of past research findings which were limited to individual or small groups of institutions. There are further reasons for doubting that there would be a connection between course packages and student performance. (i) Most research to date has assumed implicitly that students wish to maximize their mark on the examination. However, a significant proportion of students wish to maximize their probability of passing the course rather than achieving as high a mark as possible. This group would react to a technique which increased feedback by reducing the amount of time spent on the course. Table 3 summarizes student responses to the question: "In the course, my study goal is to Just Pass or Maximize Score (1, 2, 3, 4, 5)." In the event it was found that

there was no significant difference in the time students claimed to spend working on innovative and conventional courses.

(ii) It is possible that there are complicated interaction effects relating to the way in which different students react to the techniques. For example, pass-maximizing economics majors may make more use of the feedback techniques than other groups. Standard regression analysis is not well suited to the estimation of higher-order interaction effects; currently discriminant analysis and generalized linear interactive modelling are being applied to the data.

Despite the fact that the multiple regressions revealed little about the relative efficacy of the different teaching techniques, a simple specification of the model can show the main results which emerged for other variables. Not reproduced here because of space limitations are a number of regressions in which various measures of student comprehension and student opinion of the course are dependent variables; the independent variables are student characteristics and a set

TABLE 4—STUDENT OPINION OF THE CONTRIBUTION OF DIFFERENT TECHNIQUES; MICRO

Group	Institution	No. of Students	Lectures	Texts	Readings	Essays	TIPS	Cases	Tutorials	Other
A	1	163	33	23	1	16	—	—	23	4
B	2	182	26	12	8	14	13	11	14	2
C	2	291	21	25	3	—	15	19	15	2
D	4	616	36	21	6	—	18	15		4
E	4	338	23	19	6	11	19		19	3

Note: A dash denotes the technique was not included in the course package.

of dummy variables, each of which referred to a different institution. These institutional dummies should account for the various non-measured student attributes, and the interaction effects among the teaching techniques and student characteristics.

The variation in the coefficients between different measures of output renders a full discussion inappropriate, but three features do stand out: (i) There are significant, though not consistent, differences between institutions for some measures of output after controlling for student characteristics. However, for Total Micro Mark and Total Mark (i.e., combining micro and macro sections of the examination), there are no differences between institutions. (ii) Student opinion of the course is virtually unaffected by which institution was attended. This suggests that the variation in the course inputs embodied in the institutional dummies did not affect student opinion of the course. (iii) Females did better on the essays and worse on the multiple choice examinations than males. This suggests that altering the form of the examination may discriminate against certain groups of students.

While the multiple regressions give no indication of the contribution of different teaching techniques to comprehension, the students themselves provided a good deal of information. Table 4 shows student responses to the following question: "Allocate 100 points among the teaching methods to reflect the contribution each has made to your understanding of this course."

Clearly a considerable amount of substitution occurred between the techniques, and students tended to allocate about the same number of points to the highly labor-inten-

TABLE 5—CORRELATIONS BETWEEN THE DIFFERENT TYPES OF EXAMINATION

	Micro		Macro	
	Case	Essay	Multiple Choice	Essay
<i>Micro</i>				
Multiple Choice	.11 1254 (.60)	.18 1332 (.55)	.43 1483 (.75)	.18 1483 (.49)
Case		.23 881 (.61)	.10 1254 (.53)	.14 1032 (.55)
Essay			.19 1211 (.47)	.54 1211 (.74)
<i>Macro</i>				
Multiple Choice				.26 1483 (.57)

Notes: 1) Student correlations are followed by number of observations; 2) Numbers in parentheses are the mean estimated correlations of 100 U.K. academics; 3) All student correlations are significant at the 99 percent level.

sive tutorials as to the highly capital intensive TIPS. Students were also asked how much they liked each technique, irrespective of its contribution to learning. Lectures and TIPS topped the ranking with cases proving to be the least popular.

### I. Further Findings

Rendigs Fels (1970) postulated that the correlation between scores on a good multiple choice test and other measures of economics comprehension could be expected to range between .60 and .75. Fels' position is

TABLE 6—TEACHER CHARACTERISTICS: STUDENT AND TEACHER WEIGHTS

	Student		Teacher
	Micro	Macro	
Imparts enthusiasm	25	32	10
Knows subject matter well	8	9	13
Is well prepared for class	10	13	17
Demonstrates practical applications of course material	8	0	8
Presents material with clarity	24	32	16
Speaks clearly and distinctly	7	0	9
Uses the blackboard effectively	7	6	6
Is available outside class	0	-7	4
Is sensitive to students' level of understanding and is responsive to questions	10	7	6
Has respect for students' opinions	-5	0	4
Provides helpful comments on assignments	7	7	4
Number	1632	1064	154

similar to that of U.K. academics. Table 5 shows how wide of the mark informed opinion is on this issue.

These results suggest that different types of questions measure different aspects of economics comprehension and demonstrates that the form of a final examination may be a major determinant of student rankings.

For lecturers who wish to improve their teaching ratings, Table 6 contains data on the weights which students assign to various lecturer characteristics. These weights are derived from a multiple regression in which overall opinion of the lecturer is dependent variable and each of the listed characteristics the independent variables. The third column shows the weights which U.K. academics thought students would assign to the characteristics.

The results suggest that the two major factors determining good lecturing in students' eyes is imparting enthusiasm and presenting material with clarity, both of which are significantly underestimated by academics.

#### REFERENCES

- Fels, Rendigs, "Multiple Choice Questions in Elementary Economics," in K. G. Lumsden, ed., *Recent Research in Economics Education*, Englewood Cliffs: Prentice-Hall, 1970.
- Kelley, Allen C., "An Experiment with TIPS, a Computer Aided Instructional System for Undergraduate Education," *American Economic Review Proceedings*, May 1968, 58, 446-57.

# Modeling Multiple Outputs in Educational Production Functions

By JOHN F. CHIZMAR AND THOMAS A. ZAK\*

For some time now economists have conceptualized learning as a production process. Most educational production functions specify only one output—usually a measure of cognitive achievement. Some recent studies have recognized the multidimensional aspects of schooling, specifying outputs in both the cognitive and affective domains.

Introducing multiple outputs into the production function raises interesting theoretical and empirical issues. One issue is the manner in which various outputs are interrelated. How one models output interactions in educational production functions may be important. Are the outputs multiple products? If so, are they independently or simultaneously produced? Or are the outputs joint products?<sup>1</sup> Answers to these questions may affect empirical estimates of the educational production function(s).

This paper briefly describes how one might approach this problem. We argue on conceptual grounds that modeling outputs in the cognitive and affective domains as *joint* products is preferable. In addition, we compare the three approaches empirically to evaluate the sensitivity of educational production function parameter estimates to output specification.

## I. Modeling Multiple Outputs

The relationship among outputs should dictate the model (and estimating technique) one employs to estimate educational production functions. If cognitive and affective achievement are multiple products produced

with separate inputs and are completely independent, then one specifies a separate production function for each output. Letting  $Y_1$  and  $Y_2$  represent cognitive achievement and attitude toward economics, respectively, then:

$$(1) \quad \begin{aligned} Y_1 &= F(X_1, X_2, \dots, X_p), \\ Y_2 &= G(X_1, X_2, \dots, X_p), \end{aligned}$$

with  $X_i (i=1, \dots, p)$  inputs.

If the outputs are produced simultaneously, a simultaneous equations system is appropriate. Since the outputs are multiple products, one specifies a separate production function for each output; the system becomes

$$(2) \quad \begin{aligned} Y_1 &= F(Y_2, X_1, X_2, \dots, X_p), \\ Y_2 &= G(Y_1, X_1, X_2, \dots, X_p). \end{aligned}$$

A single equation production function (with multiple outputs) should be estimated if cognitive and affective achievement are joint products. Then, in implicit form, the function may be represented as

$$(3) \quad F(Y_1, Y_2, X_1, \dots, X_p) = 0.$$

For estimating purposes we may restrict the function to satisfy the functional form:

$$(4) \quad f(Y_1, Y_2) = g(X_1, X_2, \dots, X_p).$$

Which model most closely corresponds to the teaching/learning environment encountered in economics classrooms depends on the degree to which the outputs share the same inputs. This, in turn, depends upon the concept of "input exhaustion" introduced by Byron Brown and Daniel Saks (1980). It indicates the extent to which an increase in

\*Illinois State University, and U.S. Naval Academy and Federal Trade Commission, respectively.

<sup>1</sup>We follow the nomenclature introduced by Potluri Rao (1969) who distinguishes multiple products, where each output is produced under *separate* production processes, from joint production, where several outputs are produced from a single production process.

input applied to one output reduces the amount of that input available to produce other outputs. Complete input exhaustion occurs when outputs are independently produced (i.e., zero input sharing). Each output has a separate production process. If, however, using an input to produce one output does not reduce its availability for producing other outputs, then the input is shared and the outputs are jointly produced by a single production process. Thus, the extent of input exhaustion indicates the degree of jointness in production.

A variety of variables have been entered as inputs into economic education production functions, but typically one finds individual student characteristics (such as test scores, ability proxies, race, sex, and years of schooling) and classroom information (teacher attributes, teaching technique, and class size, for example). Most inputs enter both cognitive and affective achievement equations. If they enter one equation but not the other—the so-called zero restrictions hypothesis—then the inputs are not shared. Since they can contribute to both student understanding and attitude, this is reasonable. However, without less aggregate data that indicate the amount of each input expended to promote economic cognition as distinct from the amount spent enhancing students' attitudes toward economics (if it is even possible to separate them), one must enter the entire value for each input into both equations. Explicitly entering the full amount of each input into separate production functions implicitly assumes that cognitive and affective achievement are joint products (i.e., no input exhaustion). Joint products are the result of a single production process. Conceptually one should not estimate separate equations for each output, but rather a single production function embodying both outputs. Absent more detailed data on the allocation of inputs between educational outputs, single equation joint product estimates are appropriate. See our earlier work (1983) for a discussion of other reasons for preferring a joint product approach. The empirical importance of jointness and input exhaustion, however, is as yet unexplored. We now turn to this task.

## II. Estimation

If  $Y_1$  and  $Y_2$  are multiple products produced independently (so that equation (1) is the correct conceptualization of the production process), then ordinary least squares (*OLS*) is an appropriate estimating technique. If equation (2) is the correct conceptualization, that is, multiple products produced simultaneously, then two-stage least squares (*2SLS*) is appropriate. If outputs are jointly produced (equation (4)), then neither *OLS* nor *2SLS* is appropriate because, for both techniques, a single equation can have only one dependent variable. Instead, an interesting alternative procedure based on the works of Hrishikesh Vinod (1968, 1969, 1976) is used to estimate the parameters of equation (4). The procedure is an adaptation of Harold Hotelling's canonical correlation that yields parameter estimates in terms of the *original* variables instead of uninterpretable index numbers.

Technical development of the procedure is laid out completely in Vinod's articles and need not be presented here. However, a brief outline of its application to economic education production function estimates is given.

We specify equation (4) as a generalized Cobb-Douglas production of the following form:

$$(5) \quad \sum_{i=1}^2 \alpha_i \log Y_i = \beta_0 + \sum_{j=1}^p \beta_j \log X_j + \varepsilon,$$

where  $\varepsilon$  is the disturbance term.

If we denote  $a_{1i}$  and  $b_{1j}$  with  $i=1,2$  and  $j=1,2,\dots,p$  as the first canonical coefficients corresponding to  $\rho_1$ , the first canonical correlation, then estimators of the coefficients in equation (5) can be obtained as

$$(6) \quad \hat{\alpha}_i = \hat{a}_{1i},$$

$$(7) \quad \hat{\beta}_j = \hat{b}_{1j} \hat{\rho}_1,$$

where the "hats" indicate estimators and additional canonical correlations are ignored. This procedure comes closer to the theoretical equation (4) and takes account of jointness in production.

The techniques are applied to data collected in a large section of economic principles in the fall of 1978 at Illinois State University. The course was taught primarily using a lecture format—thus it is likely that input sharing occurred. The sample consists of 175 students who responded to survey instruments.

We specify the joint product production function by rewriting equation (5) using mnemonic notation:

$$\begin{aligned}
 (8) \quad & \alpha_1 \ln POSTT + \alpha_2 \ln POSTA = \beta_0 \\
 & + \beta_1 \ln PREA + \beta_2 \ln EXPG \\
 & + \beta_3 \ln PROF + \beta_4 \ln PRET \\
 & + \beta_5 \ln EFF + \beta_6 \ln CACT \\
 & + \beta_7 \ln AGE + \varepsilon,
 \end{aligned}$$

where *POSTT* = *TUCE* (Test of Understanding of College Economics) score at end of course; *POSTA* = attitude score at end of course; *PREA* = attitude score at beginning of course; *EXPG* = student's expected grade; *PROF* = student's evaluation of the professor given at the conclusion of the course (1 to 5 Likert scale); *PRET* = *TUCE* score at beginning of course; *CACT* = composite *ACT* score; *TIPSCORE* = total score on TIPS surveys, maximum possible score = 115 (TIPS is an acronym for Teaching Information Processing System); *EFF* = effort index = *TIPSCORE*/*CACT*; and *AGE* = student's age.

These variables are included to control for student background, prior achievement, ability, and effort. With the exception of "attitude" and "effort," the definitions of the inputs and outputs have been employed in previous estimation of economics learning models reported in the literature. See John Siegfried and Rendigs Fels (1979) for a review of this literature.

The attitude score is the sum of responses on a 14-item Likert scale instrument. It is designed to measure the student's attitude toward economics and was administered at

both the beginning (*PREA*) and end of the semester (*POSTA*).

The *EFF* is an effort index formed from the student's total score on the TIPS surveys (*TIPSCORE*) and his/her composite *ACT* score (*CACT*). The *TIPSCORE* indicates the extent to which the student works continuously and is a measure of academic achievement, and *CACT* is a proxy for learning capacity. Thus, *EFF* so defined imputes the student's achievement to his/her learning capacity.

To facilitate interpretation, marginal rates of substitution are derived from the Cobb-Douglas production function estimates and presented in Table 1.<sup>2</sup> (Tables of parameter estimates and summary statistics are available from the authors.) The estimated *MRS* of *EFF* for *CACT* ( $= MP_{cact}/MP_{eff}$ ) in the production of *POSTT* calculated using Vinod's adaptation is .351. This result indicates that less able students can compensate for deficiencies in ability with extra effort. An increase in the effort index of 1.75 points is necessary to offset a 5 point *CACT* disadvantage. Thus, a 1.0 standard deviation decrease in *CACT* can be offset by a 1.7 standard deviation increase in *EFF*. Compare this with *MRS* estimates calculated using *OLS* and *2SLS* of .516 and .541, respectively. The estimates are nearly equal and imply a 2.7 point increase in the effort index is necessary to offset a 5 point *CACT* disadvantage. A 1.0 standard deviation decrease in *CACT* can be offset by approximately a 2.6 standard deviation increase in *EFF*.

Specifying the outputs as joint products suggests that less able students can compensate for differences in ability with extra effort. The *MRS* estimates from the other multiple product specifications all indicate it is substantially more difficult to so compensate. As with the view that school inputs do not matter, economic educators must find it difficult to adopt a conclusion that signifi-

<sup>2</sup>We appear to be comparing apples and oranges. The empirical models for Vinod's adaptation and *2SLS* are not directly comparable because of the necessity to identify the latter. For this reason, *OLS* was used to estimate the *2SLS* model. The results are quite similar to those shown in Part B of the table and are available from the authors.

TABLE 1—MARGINAL RATES OF SUBSTITUTION: OUTPUT *POSTT*

	<i>PREA</i>	<i>EXPG</i>	<i>PROF</i>	<i>PRET</i>	<i>EFF</i>	<i>CACT</i>	<i>AGE</i>
<b>A. Vinod's Adaptation</b>							
<i>PREA</i>	1.000	12.575	3.781	.457	3.767	1.324	.370
<i>EXPG</i>	.079	1.000	.300	.036	.299	.105	.029
<i>PROF</i>	.264	3.327	1.000	.121	.997	.350	.098
<i>PRET</i>	2.190	27.540	8.280	1.000	8.250	2.900	.810
<i>EFF</i>	.265	3.388	1.004	.121	1.000	.351	.098
<i>CACT</i>	.755	9.496	2.855	.345	2.845	1.000	.279
<i>AGE</i>	2.704	34.001	10.222	1.235	10.185	3.580	1.000
<b>B. OLS</b>							
<i>PREA</i>	1.000	84.496	4.740	8.541	48.676	25.111	4.242
<i>EXPG</i>	.011	1.000	.056	.101	.576	.297	.050
<i>PROF</i>	.211	17.829	1.000	.056	10.271	5.299	.895
<i>PRET</i>	.117	9.983	.055	1.000	5.700	2.940	.497
<i>EFF</i>	.021	1.736	.974	.176	1.000	.516	.087
<i>AGE</i>	.236	19.920	1.118	2.014	11.475	5.920	1.000
<b>C. 2SLS</b>							
<i>POSTA</i>	—	—	—	1.522	8.020	4.376	.711
<i>EXPG</i>	—	—	—	—	—	—	—
<i>PROF</i>	—	—	—	—	—	—	—
<i>PRET</i>	—	—	—	1.000	5.258	2.870	.466
<i>EFF</i>	—	—	—	.190	1.000	.546	.089
<i>CACT</i>	—	—	—	.348	1.834	1.000	.163
<i>AGE</i>	—	—	—	2.147	11.312	6.173	1.000

cant tradeoffs in the learning process almost never happen. We have all had our enthusiasm for teaching buoyed by students who make up in hard work and extra effort what they lack in raw ability.

Similar conclusions emerge from comparisons of the *MRS* of *EFF* for *PRET* estimates (a measure of the tradeoff between prior economic knowledge and effort). The joint product estimate of *MRS* is .121. It implies that an increase in the effort index of .4 points is necessary to offset a 3.3 point *PRET* disadvantage. A 1.0 standard deviation decrease in *PRET* can be offset by .39 standard deviation increase in *EFF*. The *OLS* and *2SLS* estimates of .176 and .190 (again roughly equal) imply that a 1.0 standard deviation in *PRET* can be offset by (approximately) a .6 standard deviation increase in *EFF*. Thus, not only is it easier to offset prior economic knowledge than ability with extra effort, but both tradeoffs are easier when the outputs are specified as joint products. Additional tradeoffs within the learning process can be examined by investigating other entries in Table 1.

### III. Conclusion

This paper has investigated some of the theoretical and empirical implications of introducing multiple outputs into economic education production functions. We argue, using a simple verbal paradigm, that modeling outputs in the cognitive and affective domains as joint products in contradistinction to multiple products is conceptually superior. We also compare three empirical approaches—Vinod's adaptation, *OLS*, and *2SLS*—to investigate if how one models multiple outputs plays an important role in economic education production function parameter estimates and consequent *MRS*. While all three approaches suggest that there are tradeoffs within the learning process, substitution appears less difficult when outputs are modeled as joint products.

### REFERENCES

- Brown, Byron and Saks, Daniel, "Production Technologies and Resource Allocation within Classrooms and Schools: Theory

- and Measurement," in R. Dreeben and J. A. Thomas, eds., *The Analysis of Educational Productivity*, Vol. 1, Cambridge: Ballinger, 1980.
- Chizmar, John F. and McCarney, Bernard J., "Canonical Estimation of Joint Educational Production Functions: An Evaluation of a 'Trade-Offs' Implementation," *Journal of Economic Education*, 1983, forthcoming.
- \_\_\_\_\_ and Zak, Thomas, "Canonical Estimation of Joint Educational Production Functions," *Economics of Education Review*, Spring 1983, forthcoming.
- Rao, Potluri, "A Note on Econometrics of Joint Production," *Econometrica*, October 1969, 37, 737-38.
- Siegfried, John and Fels, Rendigs, "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, 17, 923-69.
- Vinod, Hrishikesh, "Econometrics of Joint Production," *Econometrica*, April 1968, 36, 322-36.
- \_\_\_\_\_, "Econometrics of Joint Production—A Reply," *Econometrica*, October 1969, 37, 739-40.
- \_\_\_\_\_, "Canonical Ridge and Econometrics of Joint Production," *Journal of Econometrics*, May 1976, 4, 147-66.

# Who Maximizes What? A Study in Student Time Allocation

By ROBERT M. SCHMIDT\*

Over the last decade the economic education literature has highlighted the evaluation of various instructional techniques. The framework has been the educational production function, positing measured course output as a function of student ability, time, various attributes, and participation in an experimental treatment. Parameters have been estimated primarily through single equation, ordinary least squares (*OLS*) regression.

More recently, the utility-maximization approach to student behavior has directed attention toward student time allocation among courses and leisure. Unfortunately, the student time variable, plausibly the most "controllable" of the various production function inputs, typically has been omitted from empirical studies. Moreover, the rare study that controls for student time does so by measuring *total* time devoted to the course, and this variable is frequently found to be insignificant. A possible explanation for this result is that the intensity of study varies so much among students that the assumption of time homogeneity is violated. An alternative explanation is that the time variable is overly aggregated. Students not only allocate scarce time among courses and leisure, they also ration time among alternative study modes within a course. The present study examines these hypotheses by evaluating the student time variable in a model with five time measures—hours spent in lectures, discussion sections, study outside of class, preparation for a midterm examination and preparation for the final examination.

Three alternative estimation techniques are contrasted—*OLS*, full-information maximum likelihood (*FIML*) for a simultaneous system, and *FIML* for a simultaneous system in which time and ability are treated as "unobservable" or "latent" variables. The latter model introduces linear structural relation-

ships (*LISREL*) into the economic education literature, permitting the investigation of multiple indicators of latent measures (dependent and independent variables) in an educational production function. The *LISREL* provides a promising approach to coping with the unobservables pervasive in educational research.

## I. The Experiment and the Data

The data used in this analysis were collected by Allen C. Kelley for the evaluation of his Teaching Information Processing System (TIPS), although they have never appeared in the literature in that regard. The TIPS utilizes the computer to manage instruction in the large-lecture setting. As used by Kelley, TIPS processes weekly, 15–20 question multiple choice surveys of student knowledge. The surveys are optional and do not count toward the course grade. Rather, they provide feedback on student progress to the students, discussion section leaders, and professor. This is accomplished through a series of reports printed by the computer and available the day after the survey. Students receive detailed information and possible remedial or enrichment assignments. Section leaders and the professor receive summary reports.

The experiment was run in the fall of 1970 at the University of Wisconsin-Madison in the macroeconomic principles course using a control and experimental lecture. While an evaluation of student time allocation is the motivation for the current study, TIPS will be analyzed as the experimental teaching technique. The data set is attractive because it is large (216 students with the necessary variables), clean, and provides detailed information concerning student time usage.

## II. Disaggregating Student Time

This section analyzes the productivity of student time in the traditional manner, that

\*Assistant professor, University of Richmond.

is, by estimating an educational production function via *OLS* regression. A Cobb-Douglas production function is employed for two reasons. First, Elisabeth Allison, and John Chizmar and Thomas Zak have found an additional transcendental component either to be insignificant, or to not alter marginal rates of substitution. Second, the logarithmic transformation is linear, simplifying the simultaneous estimation of the next section. The output of the production function is the percentage score on those multiple choice questions of the final exam (25 *TUCE* questions plus 21 of the professor's own) covering material taught during the TIPS experiment (initiated after the first examination). Inputs into the production function are: 1) a vector of proxy measures for ability, including percentile rank on the *ACT* or *SAT* exam (*ACTSAT*), high school percentile (*HSP*), first exam score (*EXAM1*), and *PRETUCE* score; 2) a vector of student hours devoted during the TIPS experiment to lecture (*LECTHRS*), discussion sections (*SECTHRS*), study outside of class (*STUDYHRS*), second exam preparation (*HRS2ND*), and final exam preparation (*HRSFINAL*); 3) a vector of binary "taste" measures including sex (*MALE*), class (*SOPH* or *UPPER*), and major (*pre-BUSINESS* or *ENGINEER*, which includes mathematics and physical sciences as well); and 4) a binary variable for the TIPS experimental lecture.

Coefficients for ability and time represent output elasticities, all hypothesized to be positive. Earlier studies have shown TIPS to exert a positive influence on student achievement and therefore its shift parameter is predicted to exceed unity. I remain agnostic on the shift parameters for the taste measures. These may exceed unity, indicating higher marginal and average products for all ability and time variables, or fall below unity, indicating lower productivity.

The production function is estimated twice, once with an aggregate time measure and once with the five separate measures. The results for the total time measure are presented as column 1 in Table 1. As in those few studies that examine student time, I find the time coefficient to be insignificant. This may be the result of excessive aggregation.

Students operating at the margin are likely to view some time uses as substitutes, others as complements. For example, students who study consistently and attend lectures religiously may avoid cramming for exams. While many students might devote approximately the same amount of time to a course, they may allocate that time among pursuits with varying productivities. The aggregate time variable may therefore mask the interesting and relevant production function responses.

This possibility appears to be corroborated by the second column of Table 1. The time variables perform largely as expected. Three are positive and significant determinants of course output. The marginal products, evaluated at the means, of the hours spent in lectures, in discussion sections, outside of class, studying for the second exam, and studying for the final exam are 1.01, 0.78, 0.05, 0.61, and  $-0.35$ , respectively. The anomaly is *HRSFINAL* which displays a negative coefficient and a *t*-value of  $-1.95$ . A negative marginal product for final exam study might be possible for the very weak student, but is questionable on average. This result could be spurious if some students who know the material well, study less and still perform well. A student attempting to recoup a lost semester, however, might study extensively and still perform poorly. While this issue cannot be analyzed in detail in the present study, a side experiment indicates that cramming might detract from final exam performance for some students. Separate regressions reveal a significant negative coefficient for weak students (*ACTSAT* scores more than one standard deviation below the mean), an insignificant negative coefficient for average students (within one deviation of the mean), and an insignificant positive coefficient for strong students (more than one deviation above the mean).

Three of the four ability variables are of the predicted sign and significant, implying that each measures something slightly different. Perhaps *ACTSAT* measures general learning ability, *EXAM1* measures a grasp of the basic tools, and *PRETUCE* measures economic knowledge and reasoning prior to the first college economics course. Also, since

TABLE 1—PARAMETER ESTIMATES USING ALTERNATIVE MEASURES OF STUDENT TIME AND ALTERNATIVE ESTIMATION TECHNIQUES

	OLS		Observed Approach			Latent Approach		
	Output (1)	Output (2)	Output (3)	<i>TIMESEM</i> (4)	<i>TIMEXAM</i> (5)	Output (6)	<i>TIMESEM</i> (7)	<i>TIMEXAM</i> (8)
<i>R</i> <sup>2</sup>	0.386	0.442	0.389	0.460	0.769	0.432	0.299	0.892
Total Time	0.045							
<i>TIMESEM</i>			0.098 <sup>b</sup>		−0.103	0.258 <sup>a</sup>		−0.338
<i>LECTHRS</i>		0.215 <sup>a</sup>				0.348 <sup>b</sup>		
<i>SECTHRS</i>		0.046 <sup>a</sup>				0.933 <sup>b</sup>		
<i>STUDYHRS</i>		0.017				0.480 <sup>b</sup>		
<i>TIMEXAM</i>			−0.002	0.034		0.017	−0.061 <sup>a</sup>	
<i>HRS2ND</i>		0.054 <sup>a</sup>				0.609 <sup>b</sup>		
<i>HRSFINAL</i>		−0.048				0.655 <sup>b</sup>		
<i>ABILITY</i>						0.158 <sup>b</sup>	0.140 <sup>b</sup>	−0.096
<i>ACTSAT</i>	0.076 <sup>b</sup>	0.058 <sup>b</sup>	0.073 <sup>b</sup>	0.010	−0.045	1.000		
<i>HSP</i>	−0.050	−0.060	−0.041	−0.057	−0.060	0.246 <sup>b</sup>		
<i>EXAM1</i>	0.156 <sup>b</sup>	0.112 <sup>b</sup>	0.157 <sup>b</sup>	−0.088	−0.221	0.165 <sup>b</sup>		
<i>PRETUCE</i>	0.195 <sup>b</sup>	0.220 <sup>b</sup>	0.198 <sup>b</sup>	0.005	0.080	0.286 <sup>b</sup>		
<i>TIPS</i>	1.076 <sup>b</sup>	1.075 <sup>b</sup>	1.072 <sup>b</sup>	1.022	0.895 <sup>b</sup>	1.074 <sup>b</sup>	1.009	0.836 <sup>b</sup>
<i>SOPH</i>	0.977	0.996	0.986	0.949	1.042	0.990	0.926	1.016
<i>UPPER</i>	1.015	1.040	1.023	0.894	0.930	1.045	0.868	0.871
<i>MALE</i>	1.021	1.034	1.024	0.938	0.942	1.083 <sup>a</sup>	0.880 <sup>a</sup>	0.914
<i>BUSINESS</i>	0.981 <sup>*</sup>	0.971	0.977	1.045	1.129 <sup>b</sup>	0.971	1.055	1.201 <sup>a</sup>
<i>ENGINEER</i>	0.966	0.962	0.959			0.944		
<i>CREDITS</i>				−0.005	−0.080		−0.015	−0.132
<i>AVEWEEK</i>				0.421 <sup>b</sup>			0.204 <sup>b</sup>	
<i>AVEXAM</i>					0.287 <sup>b</sup>			0.472 <sup>b</sup>
<i>AVEFINAL</i>					0.614 <sup>b</sup>			1.009 <sup>b</sup>

Notes: Values in parentheses are estimates of the elasticity of an observed indicator with respect to its latent variable. Constants were estimated for columns 1–2 but are not shown here. Columns 3–8 were estimated from covariance matrices, hence, constant terms are inapplicable.

<sup>a</sup>Significantly different from zero for continuous measures or unity for binaries, 95 percent level.

<sup>b</sup>Significant at 99 percent level.

high school economics courses were rare in the late 1960's, *PRETUCE* may reflect a self-motivating interest in the subject as well as a specific aptitude for it.

With respect to the binary measures, *TIPS* is the only significant variable as it adds 7.5 percent to student scores—5 percentage points at the class average of 68 percent correct. The insignificant results for the other binaries imply, for example, that time spent by males is no more productive than for their female counterparts.

### III. Reaggregating the Time Measures

Simultaneity may exist in educational research for two reasons. First, the alternative time inputs are interdependent within a student's maximizing scheme. Second, an unob-

served variable—knowledge prior to studying for the final exam—influences both the time spent studying for the final as well as the score on the final. While simultaneous estimation is appropriate in this case, the system is unwieldy. The large number of endogenous variables (final exam score and five time measures) make identification and analysis difficult. As a result, the time variables are reaggregated into two measures—time spent on a consistent basis throughout the term (*LECTHRS*, *SECTHRS*, *STUDYHRS*) and time spent preparing for examinations (*HRS2ND* and *HRSFINAL*).

Therefore, two new equations (*TIMESEM* and *TIMEXAM*) are introduced along with four new exogenous variables. The new variables influence time, but not exam score. The

first, an opportunity cost measure, is the number of credit hours taken (*CREDITS*). The other three, taste variables for studying, are *AVEWEEK*, *AVEXAM*, and *AVEFINAL*. They correspond to the time spent in an average course in an average week, in preparing for a midterm exam, and in preparing for a final exam. The remaining parameters in the *TIMESEM* and *TIMEXAM* equations must therefore be interpreted relative to time spent in the average course. For example, prebusiness majors may devote more time to economics than to other courses, in which case the estimated parameter would exceed unity. The Cobb-Douglas function is also imposed on these two equations.

The method for aggregating the time alternatives remains to be determined. Two approaches are considered. The first, the "observed" approach, simply sums the relevant hours to form a total. The alternative, the "latent" approach, treats *TIMESEM* and *TIMEXAM* as unobservable in the sense that each reflects a productivity-adjusted time input. Thus, not only are measured output and *TIMEXAM* functions of *TIMESEM*, but so are *TIMESEM*'s observable "indicators," *LECTHRS*, *SECTHRS*, and *STUDYHRS*. A parallel argument applies to the relationship between the unobservable *TIMEXAM* and its observable indicators, *HRS2ND* and *HRSFINAL*. A similar approach might be taken with respect to student ability. The ability to learn economics is truly unobservable, but with observed indicators *ACTSAT*, *HSP*, *EXAM1*, and *PRETUCE*. The "measurement" model for the unobservables is thus comprised of nine equations of the Cobb-Douglas format. Five pertain to the endogenous unobservables, and four to the exogenous *ABILITY* unobservable.

The *LISREL* computer package was employed in the estimation of both the observed and latent variants of the model. The *LISREL* applies the full-information maximum likelihood (*FIML*) method to the estimation of simultaneous systems. When the system includes unobservables, *LISREL* incorporates the measurement model into the simultaneous equations model. An alternative technique is to estimate separate

indexes for the unobservables via factor analysis and then use these indexes in the simultaneous estimation. This approach ignores the information contained in the structural relationships when estimating the indexes.

The estimates are presented in Table 1. Consider first the results when using observed time and ability measures (cols. 3–5). The output equation provides results which are similar to those of *OLS*. Parameter estimates are of comparable magnitudes and significance. The exceptions are the time aggregates, *TIMESEM* and *TIMEXAM*, which fall between the *OLS* estimates for their component measures. The *TIMESEM* is significant, while *TIMEXAM* is insignificant.

Very little appears to affect the total amount of time students spend on this economics course beyond their taste for studying (*AVEWEEK*, *AVEXAM*, *AVEFINAL*). Students approach this as an average course with two exceptions. Prebusiness majors spend about 20 percent more time studying for exams in this course. The output equation indicates, however, that they have no advantage in learning economics. The second exception is for students in the TIPS lecture. The TIPS students spend no more time during the semester, but about 10 percent less time studying for exams. Apparently the weekly surveys do not induce students to spend more time during the semester. Rather the surveys focus study time as the 7 percent higher output in the TIPS class would indicate.

Consider next the estimates of columns 6–8, the unobservables approach. The first issue which arises is the comparability of results between the observable and unobservable estimates. Comparability has been ensured by constraining the error variances of the latent *TIMESEM* and *TIMEXAM* equations to the same values estimated for their observed counterparts. This defines equivalent scales for the observed and latent measures. Given this, the most striking contrast between the two estimates lies in the increase of the *TIMESEM* output elasticity for 0.098 in the observed estimation to 0.258 in the latent. The explanation likely lies in the alternative weighting schemes. The ob-

served approach weights equally time inputs which may exhibit unequal productivities. The latent estimation relaxes that assumption. The estimated weights are noted in parentheses and can be interpreted as elasticities. Thus, a 1 percent increase in *TIMESEM* is equivalent to a 0.348 percent increase in *LECTHRS*, a 0.933 percent increase in *SECTHRS*, or a 0.048 percent increase in *STUDYHRS*. The implication is that hours spent in lecture are the most productive, hours spent studying are slightly less productive, and hours spent in discussion are substantially less productive.

The *TIMEAM* can be analyzed analogously. The observed indicators are weighted almost equally. The *TIMEAM*'s coefficient in the output equation remains insignificant, although it is now of the predicted sign. The more interesting change is in the *TIMESEM* equation. Where *TIMEAM* had been positive and insignificant in the observed estimation, it is now negative and significant at the 95 percent level. This result accords with the hypothesis that cramming is a substitute for study during the semester.

The latent *ABILITY* variable cannot be interpreted as cleanly as have the two time measures. Whereas the time components can be summed to form a comparable observed variable, the proxy measures for ability cannot. Consequently, the scale for *ABILITY* is set arbitrarily by constraining *ACTSAT*'s weighting parameter to unity. While the interpretation of *ABILITY* is abstract, its treatment as unobservable brings about two changes. First, *ABILITY* is now significant in the *TIMESEM* equation while none of its proxies are significant in column 4. Apparently the savvy student can discern the more productive time uses while the less able student cannot. Second, the *MALE* binary is now significant indicating that the male student's productivity is 8.3 percent higher than the female's, and that males study 12 percent less during the semester. This turnabout emphasizes the role of collinearity in educational research. Apparently, the *MALE* binary is collinear with one or more indicators for *ABILITY*. Inclusion of several correlated proxy variables for an unobservable can mask the significance of other relationships.

#### IV. Conclusions

This study has extended the student utility-maximizing framework to examine student time allocation within a course. I find that when treated as an aggregate measure, time spent in an economics course does not alter student learning. On the other hand, components of that time measure do play significant roles. Hours spent in lectures, in discussion sections, and in studying for the second examination are all positive and significant. Studying for the final examination may have a negative marginal product for weak students.

A simultaneous equation system was formulated with measured output, hours spent during the semester, and hours spent preparing for exams as the endogenous variables. This system was estimated using observed time measures and then again using an unobservables approach for the two time measures as well as for student ability. Either approach provides the same conclusions for this study's experimental technique, TIPS. The TIPS students spend no more total time during the semester, but they study less for exams and exhibit higher time productivities than do students in the control lecture. The unobservables approach clarifies the analysis in several other respects. The latent time variables demonstrate higher output elasticities. Moreover, two of the latent variables enter equations significantly whereas their observed corollaries do not. More importantly, by reducing the number of highly correlated variables, the unobservables approach can reduce the collinearity which confounds educational research. In the present study, the learning advantage of male students is apparent in the unobservables, but not observables estimation.

#### REFERENCES

- Allison, Elisabeth, "Educational Production Function for an Introductory Economics Course," in Rendigs Fels and John J. Siegfried, eds., *Research on Teaching College Economics*, New York: Joint Council on Economic Education, 1982.

Chizmar, John F. and Zak, Thomas A., "Canonical Estimation of Joint Educational Production Functions," *Economics of Education Review*, Spring 1983, forthcoming.

Joreskog, Karl G. and Sorbom, Dag, *LISREL: Analysis of Linear Structural Relationships*

*by the Method of Maximum Likelihood, User's Guide*, Sweden: University of Uppsala, 1981.

Kelley, Allen C., "The Student as a Utility Maximizer," *Journal of Economic Education*, Spring 1975, 6, 82-92.

## ECONOMICS OF FERTILITY

### Mortality Rates, Mortality Events, and the Number of Births

By RANDALL J. OLSEN\*

Demographic transition theory views a decline in infant mortality as a precondition for a decline in fertility. This paper describes some new methods that can be applied to cross-sectional data on families to investigate a central issue in transition theory—the replacement hypothesis.

The term replacement is often used to describe the process by which higher mortality is translated into higher fertility, although there is a variety of channels through which replacement may run. First, there is direct replacement. This describes a conscious action by a couple to increase the number of children born in response to the actual death of one of their children. In order for direct replacement to exist, the couple must have preferences over the number of children that survive, and the will and ability to alter the timing and number of births in order to move toward their fertility goal.

The second channel of replacement is hoarding. Whereas direct replacement refers to actions taken in response to actual deaths in furtherance of a couple's fertility goals, hoarding refers to actions taken in response to anticipated deaths. Hoarding involves differential actions by couples responding to the different underlying mortality rates they face. A third replacement channel related to hoarding is societal replacement, which refers to customs of a culture that arise in response to a common level of mortality. For example, taboos against intercourse during religious festivals may be practiced to enable the society to attain a level of fertility which generates a reasonably stable population. In-

sofar as differential societal replacement exists for the Malaysian villages in my data, it may emerge as hoarding in my estimates.

Fourth, there is biological replacement which arises because, for physiological reasons, the death of a child shortens the interval to the next birth. If breastfeeding prolongs the period of sterility after birth, then when a nursling dies, this period of sterility will be shortened, mimicking direct replacement.

The objective of this paper is to show how the effects of direct replacement, hoarding, and biological replacement can be separated and then estimated. The focus will be on the methodology, although data from the Malaysian Family Life Survey will be used to illustrate the method. The methods employed use statistical techniques more fully described elsewhere (see my paper with Kenneth Wolpin). The final estimate for the rate of replacement in the Malaysian data is somewhere from 30 to 40 percent, with the biological effect via breastfeeding accounting only for about 12 percent. Replacement is greatest for children who die soon after birth with the attempt to replace being concentrated early in the birth interval. This timing of replacement makes it difficult to separate from the effects of breastfeeding. There is also an indication in the data that breastfeeding is used as a means of contraception.

#### I. Family Size Regressions

Because the central issue in the replacement literature is the effect of the death of a child on fertility, it would seem the most direct strategy would be to regress the number of births in a family on the number of infant deaths.

However, the simple regression of births on deaths will produce a misleading estimate

\*Department of economics, Ohio State University. I would like to thank the Rockefeller and Ford Foundations whose generous grant made this research possible. T. Paul Schultz and Susan Watkins made very helpful suggestions on a more detailed version of this paper. That version is available upon request.

of direct replacement. Even if families do not follow a replacement strategy, families with more births will tend to have more deaths simply because they have more children at risk. This will produce a positive coefficient on deaths unrelated to replacement. The estimation of the rate of replacement by regressing births on deaths is discussed in my 1980 paper and my paper with James Trussell.

In my 1980 paper, I showed how one can correct the least squares coefficient of deaths so as to remove the bias caused by these two factors. The method was refined in my paper with Trussell to allow for the possibility of variation in the rate of replacement across families.

In order to estimate the fertility hoarding component of replacement in cross-sectional data, it will be necessary to relate variations in fertility across families to variations in the child mortality rate across families. The true child mortality rate for a family is not observable. We can observe only the realized child mortality rate for a family, which measures the true rate with error. The child mortality rates for families may differ because of actions taken by the family. For example, if a family allocates more parental time to child care, one would expect it would have a lower mortality rate. If conscious actions concerning inputs of time to child care are correlated with conscious actions to have children, the family's observed mortality rate may be related to fertility not because of hoarding, but because parents who desire more children also like to spend more time with them and so suffer a lower rate of child mortality. To avoid this source of contamination, it is necessary to calculate the family mortality rate net of those factors which affect child survival and are possibly subject to parental choice. To do this I estimate a model of the waiting time to the death of a child over the first ten years of life for each child in the family. The method of estimating this waiting time model is described in my paper with Wolpin. The method involves the estimation of a regression equation with the length of life of a child as the dependent variable. The regressors include variables describing the physical surroundings of the household (sanitation,

source of drinking water, etc.), variables measuring the allocation of time to child care, the number and ages of other siblings, sex and birth weight of the child, breastfeeding, and finally a family specific fixed effect. This last variable captures variations in the child mortality rate which are due to factors which do not change during the time the family's children are being raised, such as the backgrounds of the parents. If all the inputs of time and goods which contribute to child survival are included, the fixed effect captures the cross-family variation in biological (and ecological) factors influencing child survival. So long as these factors are known to the family, their effect on fertility should reveal the effect of exogenous changes in the child survival rate.

Once the family specific component of the child mortality rate has been estimated, one can regress the number of births on the number of deaths and the family specific mortality rate. After correcting for the non-behavioral covariance between births and deaths, the coefficient on deaths will estimate direct replacement and the coefficient on the family specific mortality rate will estimate the hoarding response to anticipated deaths. For the Malaysian data, this procedure produces an estimate of direct replacement of 0.17 and a rate of replacement due to hoarding of 0.14. The estimated family specific mortality rate measures the true rate with error, and the anticipated mortality rate to which couples respond is probably different from the true rate, so the mortality rate regressor measures the subjective rate of mortality with error. The estimated hoarding effect is therefore biased towards zero and the direct replacement effect will also be biased although in an uncertain direction.

## II. Conception Interval Analysis

The family size regressions above cannot reveal the effects of breastfeeding, nor can they indicate the speed with which couples replace dead children. Because the Malaysian data used here gives detailed information on the dates of birth, weaning, and death of children, it is possible to directly observe the effects of breastfeeding or the death of a child on the probability that

another conception leading to a live birth results. The best way to exploit this data is to estimate a waiting time model which seeks to explain the time from the birth of a child (or the marriage of the woman) until the conception leading to the next live birth. The model developed in my paper with Wolpin is again used to estimate the conception interval waiting time model. Some of the regressors used are unchanged during the interval, such as attributes of the couple and the number of deaths which occurred before the start of the interval. Other important regressors change value during the interval, such as whether the mother is currently breastfeeding an infant, or whether a death has occurred sometime between the start of the interval and the current time. The use of explanatory variables whose values change over time substantially complicates the estimation of waiting time models, but this complication is unavoidable if we hope to separate the effects of breastfeeding and direct replacement. As noted earlier, when a nursing dies, lactation ceases, so if the effects of either breastfeeding or direct replacement on the lengths of birth intervals are to be accurately estimated, both effects must be accurately estimated.

Insofar as breastfeeding produces a replacement type response, it is possible parents exploit this biological phenomenon as a part of their strategy of direct replacement. If this is so, part of the biological breastfeeding effect must be counted as behavioral direct replacement (see T. Paul Schultz, 1976).

It should be pointed out that the waiting time from marriage to the first conception is important information, since this interval reflects fecundability in the absence of breastfeeding. Because most mothers in Malaysia breastfeed their children for at least a few months, the interval to the first birth provides the best information about fecundability in the absence of breastfeeding. An important question is whether couples have lower fecundability immediately following marriage since the woman is likely to be young. A comparatively long first interval might reflect adolescent subfecundity rather than a small ability of breastfeeding to lengthen birth intervals.

The well-known rapid rise in natality rates moving from women under 20 to women

20–24 is either eliminated or reversed when one looks at age specific natality for married women (see United Nations, 1976). The sharpness of the rise in certain developed countries suggests this may be due to pregnancy inducing marriage. Such induced marriages reduce the mean interval from the date of marriage to the first birth. This will lead to an overstatement of the ability of lactation to reduce birth intervals and, concomitantly, to an understatement of the effect of direct replacement. To prevent extremely young women from entering the sample, no first interval was begun before age 15.

### III. Results

Three conception interval waiting time specifications were used. The first has a constant term and variables for deaths before the interval and deaths during the interval. The second adds the breastfeeding variable along with some other variables, and the third includes fixed effects. The inclusion of fixed effects controls for any couple-specific fecundity effects such as desire or ability to have children (including any fertility hoarding effect.) Because the fixed effect method controls for all the time invariant characteristics of the couple, it will yield the most reliable estimates of replacement.

The results for the various replacement specifications are given in Table 1. When the

TABLE 1—ESTIMATES OF REPLACEMENT

Method	Breast-feeding Effect	Behavioral Replacement	Total
<i>Regression of Births on Deaths</i>	— <sup>a</sup>	0.21	0.21
<i>Regression of Births on Deaths and Mortality Rate</i>	— <sup>a</sup>	0.31	0.31 <sup>b</sup>
<i>Waiting Time</i>			
Mortality variables only	— <sup>a</sup>	0.90	1.04 <sup>c</sup>
Mortality and other regressors	0.18	0.75	1.07 <sup>c</sup>
Add fixed effect to above	0.15	0.17	0.46 <sup>c</sup>

<sup>a</sup>Specification cannot separate breastfeeding from direct replacement.

<sup>b</sup>Hoarding effect is 0.14, direct replacement is 0.17.

<sup>c</sup>Includes a hoarding effect of 0.14 which was estimated in second equation.

number of births is regressed on the number of deaths and corrected for spurious correlation, the estimated replacement rate is 0.21. The uncorrected least squares replacement rate is 1.4 which demonstrates the large effect of the spurious correlation.

The first waiting time regression, which has mortality variables only, estimates the replacement rate at 0.90 with over two-thirds of that resulting from an earlier conception for the interval in which the death occurs, and the rest due to the death shortening subsequent intervals.

When breastfeeding and other regressors are added, the rate of replacement is 0.93 with half occurring within the interval of the death, a third in subsequent intervals, and the rest of replacement being due to shortened breastfeeding.

Finally, with the fixed effects included the rate of replacement (excluding the effects of hoarding) shrinks to 0.32 with half the effect occurring in the interval of the death and the other half due to shortened breastfeeding. The smaller estimated effects of deaths and breastfeeding in the fixed effects regression reveal the correlations between deaths, breastfeeding, and the fixed effects. In particular, it appears that if we interpret the fixed effect as revealing desired fertility, those couples wanting larger families suffer higher mortality rates and breastfeed their children less. A positive correlation of mortality rates and fertility was also observed in detailed

analysis of the corrected least squares regression not only for this data but also for Colombian data. The relation between the fixed effect and breastfeeding suggests that families desiring fewer children are more likely to breastfeed, suggesting breastfeeding is being used as a contraceptive. If this is true, part of the breastfeeding component of replacement should be counted as behavioral.

## REFERENCES

- Olsen, Randall J., "Estimating the Effect of Child Mortality on the Number of Births," *Demography*, November 1980, 17, 429-43.
- \_\_\_\_\_ and Wolpin, Kenneth I., "The Impact of Exogenous Child Mortality on Fertility: A Waiting Time Regression with Dynamic Regressors," *Econometrica*, March 1983, 51.
- Schultz, T. Paul, "Interrelationships Between Mortality and Fertility," in Ronald G. Ridker, ed., *Population and Development: The Search for Selective Interventions*, Baltimore: John Hopkins University Press, 1976.
- Trussell, James and Olsen, Randall J., "Evaluation of the Olsen Technique for Estimating the Fertility Response to Child Mortality," mimeo., Princeton University, 1981.
- United Nations, *Demographic Yearbook*, 1975, New York: United Nations, 1976.

# Economic Analyses of the Spacing of Births

By JOHN L. NEWMAN\*

In the last several years, economists and demographers have been increasingly interested in the spacing of births. This interest has arisen partly because younger women are often observed to play an important role in the decline of fertility in a population. Their behavior cannot be analyzed adequately within a static model of completed family size, because they cannot be assumed to have completed their fertility. The observations on these young women are censored by the survey date and reflect decisions on the spacing of births as well as the desired number of births. Thus, even if one had no interest in the spacing of births, *per se*, an analysis of spacing is necessary to detect whether or not changes in fertility behavior at younger ages will be reflected some years later in completed family size.

The spacing of births can be interesting for its own sake. From an empirical standpoint, spacing can affect the rate of growth of a population through its effect on the average time between generations. There is also some evidence that the spacing affects the health of the child in its early years. Moreover, one might expect that as the variance of completed family size decreases in developed countries, many of the behavioral responses to socioeconomic conditions may come in the timing of births.

The effort spent on developing dynamic models of fertility can also be justified on theoretical grounds because some questions cannot be handled within a static framework. How the effect of explanatory variables on fertility changes at different ages and how the spacing decision interacts with female labor force participation are two important questions which come readily to mind.

Finally, there is a purely pragmatic reason for the interest. Many more data sets containing complete fertility histories are being made available from a variety of sources, notably the World Fertility Survey. This information on the timing of births represents an increase in information over data sets that contain only the number of births. It may or may not prove helpful in our understanding of fertility behavior and the reasons for fertility declines. At present, there are grounds for optimism.

This paper discusses some of the theoretical and empirical approaches that economists have taken in analyzing the spacing of births, with the hope that it will provide some guidance to the new researcher eager to exploit the richer data sets that are now available. Due to space considerations, only the economic literature employing individual data will be discussed, and the demographic and sociological literatures will not be considered.

## I. Theoretical Approaches

The theoretical approaches to the spacing of births can be divided into two broad classes. The first class maintains a static decision-making framework, but adds a spacing dimension to the decisions parents make. The simple stock adjustment models (T. Paul Schultz, 1976; Ronald Lee, 1980) assume that some proportion of the divergence between actual fertility and desired completed family size is reduced in a given time period. Although the speed of adjustment is allowed to vary over time in response to declining fecundity or general economic conditions, it is not chosen by the family as an outcome of some maximizing behavior.

Other essentially static models do allow families to exercise some discretion over how fast they have births, and attempt to model the factors that may influence that decision.

\*Department of economics and the Murphy Institute of Political Economy, Tulane University. Research support from the Rockefeller Foundation is gratefully acknowledged.

In both Assaf Razin (1980) and my 1981 study, the number and average interval between births enters into the parents' utility function. Previous research had suggested that longer intervals are associated with better health and higher attainment of the child in later years. Accordingly, a longer birth interval is viewed by Razin as increasing the quality of the child. In the absence of perfect capital markets, longer birth intervals may also be preferred because they permit a more balanced consumption of nonchild related commodities over time.

In Razin's model, the woman is assumed to face different wages before, during, and after the childrearing period. The optimal spacing pattern is generated by a tension between the desire to space births farther apart and the costs of receiving a lower wage during the childrearing period. In my study, the tension arises from the economies of scale of having births closer together. To the extent that these models develop predictions based on their assumed determinants of spacing behavior, they represent an improvement over the stock-adjustment models. However, being static models, they cannot predict how responses to variables will change over time.

The second class of models is truly dynamic in nature. In these models, utility in different time periods is a function of some set of commodities and the number of children in that time period. Unlike the static case, where, for example, the production of children is generally assumed to be more intensive in the woman's time, there is no generally accepted set of assumptions on the production technology in the dynamic setting. Many of the theoretical models discussed below attempt to formulate a completely general model. However, I suspect that, at a minimum, assumptions on the variation in costs with the child's age and possible economies of scale will be necessary to generate predictions from a dynamic model. Indeed, I suspect that in time these assumptions will be accepted with the same equanimity as they are in the static models.

The theoretical models have been framed in either continuous or discrete time and have included either continuous or discrete

outcomes. Different estimation techniques are applied depending on the direction taken in the theoretical model, and a dominant strategy has yet to emerge.

Wim Vijverberg (1982) is concerned with the discrete nature of births and converts the problem into a choice among continuous outcomes by having parents pick the time to have a birth. This choice is made in accordance with their desired number of births, a decision which is determined outside of the model. If the desired number of births is allowed to change over the parents' lifetime, it is difficult to see what analytical advantage this two-stage approach yields over a model of the desired family size at each time period. In the former approach, a desired completed family size must be chosen at each period, while in the latter approach, a desired current family size must be chosen at each period. The problem of facing discrete outcomes is not avoided.

In most of the models, the parents choose the number of births, which are viewed as continuous by V. Joseph Hotz (1980), and Robert Moffit (1980), and discrete by Kenneth Wolpin (1982). As Wolpin notes, a discrete formulation is preferable if estimation is to follow theory directly. In all of these models, the decision to have a birth or not reduces to deciding whether some expression is greater than or less than a critical value (the bang-bang solution typical of many optimal control problems). Although the same solution principle is employed, the models generate different spacing patterns as a result of their different underlying structure.

## II. Imperfect and Costly Contraceptives

These behavioral models are set in worlds of perfect and costless contraceptives. While this assumption might be reasonable for developed countries, it is less acceptable for developing countries. Incorporating imperfect and costly contraceptives does present problems for models that employ births as the choice variables and have solutions that depend solely upon a critical value. As long as the decision remains either zero or one, to have a birth or not, the implication is that the woman should either use the best avail-

able contraceptive or none at all. This conclusion is not consistent with observations that, particularly in developing countries, women use contraceptives of varying degrees of effectiveness. Coupled with survey responses suggesting that many parents do not hold *explicit* timing goals, this may lead some to reject the notion that parents space their births, and therefore deny the relevance of a spacing model. Unfortunately, such a view does not help to solve the aforementioned problem with censored observations.

An alternative approach is to follow the spirit of James Heckman and Robert Willis (1975), and model the problem as one of choosing the risk of a birth, rather than the number of births. As will be seen later, this has a natural empirical counterpart in a hazard rate estimation procedure. In this approach, parents consider all the relevant costs including contraceptive costs, and choose a level of contraceptive efficiency between zero and 100 percent. At a zero level, women would be reproducing at the biologically maximal rate, which would decrease with age as fecundity declined.

To be applicable to both developed and developing countries, the model must not be specified in a way that rules out the choice of a contraceptive in an intermediate range of effectiveness. It may well be that if women in developed and developing countries have different attitudes towards uncertainty or face different costs, those in developed countries may choose to use either a perfect contraceptive or none at all, while those in developing countries may choose an intermediate position. Of course, any birth interval of a given expected length could be achieved with either technique, although the variance of the outcome would differ. This possibility does not suggest that a different model (or no model at all) is needed for developing countries, but rather that one explore the conditions that would lead to the different choices being made.

Unfortunately, the cost of allowing an intermediate solution is additional complexity. The control chosen by the parents (the risk of a birth) would have to enter the parents' objective function in a nonlinear fashion. In such a model it may be possible to ascertain

the effects of explanatory variables on fertility behavior at different ages only through numerical simulation. Nevertheless, this may prove more instructive in interpreting the empirical results at different ages than an examination of the decision criterion of a bang-bang solution.

### III. Empirical Approaches

Statistical work must address three major problems: the aforementioned censoring of the observations on younger women; the possible correlation between the lengths of the birth intervals; and the role of the biological process that translates the behavioral goals into the fertility outcomes.

The distribution of the number of births or the lengths of birth intervals can be limited by the incomplete observation period; however, the determinants of the probability of a birth can be estimated without giving rise to a censoring bias. This estimation is accomplished by expressing the likelihood of the set of all possible outcomes within a given time period, and then estimating the parameters that maximize the likelihood of observing the particular sample outcome. A woman's fertility history over a number of years is the likelihood of observing a particular sequence of outcomes (for example, no birth followed by a birth). Focusing on the probability of a birth not only handles the censoring problem, but also provides the natural choice for translating a theoretical model with a zero-one decision into an empirical framework.

Although Hotz, Moffit, Vijverberg, and Wolpin all share this general approach in their empirical formulation, they differ in their emphasis on other problems. Wolpin is primarily concerned with having the empirical estimation follow as closely as possible from the theoretical specification. He develops a theoretical model in a stochastic framework that can be estimated, but requires the solution of a dynamic programming problem within the maximization routine.

Moffit is concerned with the second problem mentioned above, that of the possible correlation between fertility outcomes. In his model, the decision to have a birth or not

depends on the value of a latent variable composed of socioeconomic variables and a random error term. The probability of having a birth also depends upon one's present state, which, in turn, depends on past error terms. He develops a multinomial probit model where the error terms of the latent variables are assumed to follow a multivariate normal distribution.

The third problem, that of the biological process generating births, usually receives more attention from demographers than economists, but cannot be ignored in empirical studies of the spacing of births. There are two random processes involved. One arises because the included socioeconomic variables are unlikely to predict exactly the decision to have a birth; and the other arises because, given the desire to have a birth, the interval to a birth is the outcome of a waiting-time process. Empirical studies of the economic determinants of the probability of a birth within a given time period that consider the first process but ignore the second will undoubtedly possess two problems. The normal density function assumed in the maximum likelihood estimation is unlikely to be correct. Secondly, unless the waiting-time process is exponential and ignores its past, the outcomes within the given period will depend on how long the woman has been without a birth as she enters that time period. This left-censoring problem is often present with panel data. When the time period considered is short, as in Hotz who looks at the fertility outcomes over only one year, the final estimation results may be substantially affected.

Since most retrospective demographic data sets which contain information on the timing of births provide a complete history up to the survey date, the left-censoring problem can be avoided by beginning at the beginning, the age at menarche. Choosing to start with the age at menarche, which can be assumed constant or random across the population, ensures that the observations are not conditioned on a choice variable.

Although both random processes could be included by working with mixture densities to describe the probability of a birth, the task is perhaps most easily handled with a

hazard rate approach. Hazard rate analysis is proving useful in a variety of problems within economics and sociology. In this approach, women are assumed to be continuously subject to a risk of a birth. The risk is given by the hazard rate, defined as the probability of a birth, given that no birth has yet occurred in the interval. The hazard rate may vary randomly across the population (referred to as heterogeneity) and may vary with the time spent in the birth interval (referred to as duration dependence). The risk of a birth will, of course, be partly determined by behavioral factors and the hazard rate can be allowed to vary with a family's socioeconomic characteristics. For example, an additional year of education of the female may be associated with a lower risk of a birth throughout the interval.

Thus, the hazard rate consists of two components: a systematic part, the contribution from the explanatory variables, and an unsystematic part, a random individual-specific term (the heterogeneity component). The unsystematic component allows for the first source of randomness mentioned above, that the included variables will not predict exactly the decisions. The second random process is also present, since specifying the hazard rate uniquely determines a particular waiting-time density function. If the hazard were constant over time and across the population, a simple exponential density function would result.

Briefly, the estimation procedure formulates the likelihood function describing a woman's fertility history up to a given age as the product of density functions governing the lengths of birth intervals and the probability of having no birth from her last birth before the given age to that given age. A limited form of correlation is permitted among the birth intervals. The likelihood function can be equivalently expressed in terms of hazard functions with systematic and unsystematic components. A distribution for the heterogeneity component is assumed and the unobserved individual specific term is integrated out of the likelihood function. The parameters of the hazard function are then chosen to maximize the likelihood function. The estimation procedure, the

robustness to the assumed distribution of the heterogeneity component, and some illustrative results are discussed in my paper with Charles McCulloch (1982).

#### IV. Conclusions

No result has yet emerged that is likely to influence subsequent research to the degree that the finding of a negative correlation between women's education and fertility guided research on completed family size. Nevertheless, there are some observations that suggest the line of research may be fruitful.

Wolpin and my 1981 study both found that a lower survival probability is associated with an earlier age at first birth. I also found a tendency for the survival probability to have a stronger effect on the risk of births at later years. This suggests that the response of fertility to mortality may be to extend the entire childbearing period more than altering the intervals between births.

As one would expect, higher female educational levels are associated with lower risks of a birth. In a study in Costa Rica, male educational levels were also negatively related to fertility. What is more interesting is that the effects of male and female education levels on the risk of a birth are considerably more similar during the middle of the childbearing period than at either end (see my 1981 study). A possible explanation for this is that women with high values of time may gain more from the scale economies of spacing children closer together. Vijverberg finds evidence of substantial economies of scale in rearing children.

One of the reasons for optimism expressed in the introduction is that the empirical approaches described above can use individual data to replicate the patterns of fertility behavior revealed from aggregate data. These techniques permit time-varying variables, so that the fertility behavior can be associated with the values of the socioeconomic variables that prevailed at the time the decisions were made. The proper treatment of such variables will require further refinements in both theoretical and empirical approaches.

#### REFERENCES

- Heckman, James J. and Willis, Robert J., "Estimation of a Stochastic Model of Reproduction: An Econometric Approach," in Nestor E. Terleckyj, ed., *Household Production and Consumption*, New York: Columbia University Press, 1975.
- Hotz, V. Joseph, "A Life Cycle Model of Fertility and Married Women's Labor Supply," mimeo., Carnegie-Mellon University, October 1980.
- Lee, Ronald D., "Aiming at a Moving Target: Period Fertility and Changing Reproductive Goals," *Population Studies*, July 1980, 34, 205-26.
- Moffitt, Robert, "Life Cycle Profiles of Fertility: A State Dependent Multinomial Probit Model," mimeo., Rutgers College, August 1980.
- Newman, John, "An Economic Analysis of the Spacing of Births," unpublished doctoral dissertation, Yale University, December 1981.
- \_\_\_\_\_ and McCulloch, Charles, "A Hazard Rate Approach to the Timing of Births," Murphy Institute of Political Economy Discussion Paper No. 2, Tulane University, 1982.
- Razin, Assaf, "Number, Spacing, and Quality of Children: A Microeconomic Viewpoint," in J. L. Simon and J. DaVanzo, eds., *Research in Population Economics*, Vol. 2, Greenwich: JAI Press, 1980, 279-94.
- Schultz, T. Paul, "An Economic Interpretation of the Decline in Fertility in a Rapidly Developing Country: Consequences of Development and Family Planning," in R. Easterlin, ed., *Population and Economic Change in Developing Countries*, Chicago: University of Chicago Press, 1976, 209-65.
- Vijverberg, Wim P. M., "Discrete Choices in a Continuous Time Model: Lifecycle Time Allocation and Fertility Decisions," Economic Growth Center Discussion Paper No. 396, Yale University, February 1982.
- Wolpin, Kenneth I., "An Estimable Dynamic Stochastic Model of Fertility and Child Mortality," mimeo., Yale University, August 1982.

# Consumer Demand and Household Production: The Relationship Between Fertility and Child Mortality

By MARK R. ROSENZWEIG AND T. PAUL SCHULTZ\*

Two important demographic regularities are the strong positive correlations between birth rates and infant mortality rates across countries, communities, and families at one point in time, and the roughly parallel decline over time in mortality and fertility in developed countries. Within developed countries such as the United States, moreover, certain groups appear to exhibit both high fertility and high child mortality rates (for example, blacks). To the extent that public resources are allocated to improve child health and reduce child mortality as well as to help people avert unwanted births, it should be useful to understand the forces generating these noted correlations.

A number of hypotheses, with different implications for the relative efficacy of alternative health and family planning programs, have been offered to explain the association between births and child deaths. First, the associations could reflect biological relationships. Infant deaths and fertility are biologically linked in two ways: 1) a death of a child truncates breastfeeding and thus shortens the sterile period following a birth; exposure to the risk of conception increases, therefore, when infant death rates are higher; and 2) the probability of an infant's death may be biologically affected by the number of prior births; that is, birth order or the mother's cumulative fertility may directly affect the infant's health. A second hypothesis is that fertility and household investments in infant health are choices that jointly reflect the environment of the household—prices, resource constraints, health conditions.

Third, parents' purposive fertility behavior may be responsive both to the anticipated exogenous prospects of child survival (hoarding) and to the realizations of infant deaths (replacement). In this paper we report estimates of both the biological and behavioral linkages between infant mortality and fertility in the United States, which take into account heterogeneity in health and parents' choice of health inputs.

## I. Dynamic Optimizing Behavior, Health Heterogeneity, and Health Production

Variations in child death rates across families may be due to exogenous differences in the health endowments of children or in the healthiness of residential areas as well as to differences in investments in children by parents. Clearly, simple associations between fertility and actual infant deaths cannot shed much light on how fertility responds to exogenous changes in the healthiness of an environment that are brought about, for example, by regulating pollution or eradicating malaria-carrying mosquitoes. Nor do simple associations indicate how family size affects infant mortality biologically. When prices and socioeconomic characteristics have been controlled, there remain unmeasured differences in tastes and in health endowments as confounding sources of the observed fertility-mortality associations. However, estimates can be obtained of both the biological relationship between fertility and infant mortality (fertility affecting infant survival) and the effect of exogenous changes in the health environment on fertility if data are available on (i) all of the important types of behavior affecting infant survival and (ii) prices and income constraints facing households.

Consider a family in which each child has a common exogenous family health "endow-

\*Departments of economics, University of Minnesota and Yale University, respectively. Research was supported in part by grants from NIH, Center for Population Research, HD 12172 and NSF, SOC-78-14481. We have benefited from comments received at the Labor Workshop at Columbia University.

ment"  $\mu_i$ , which contains family-specific genetic and environmental attributes affecting child health. The endowment  $\mu_i$  is assumed to be known by each family and to differ across families, and is thus one source of health heterogeneity in the population. The health of child  $j$  at birth in family  $i$ ,  $H_{ij0}$ , is given by the health production function

$$(1) \quad H_{ij0} = \Gamma(Z_{ij0}) + \mu_i + \varepsilon_{ij0},$$

where the  $Z_{ij0}$  are prenatal inputs, including birth order, and  $\varepsilon_{ij0}$  is the stochastic component of health which is observed at the birth of the child. In the next period of the child's life, the health production function is

$$(2) \quad H_{ij1} = \Gamma(Z_{ij0}, Z_{ij1}, \varepsilon_{ij0}) + \mu_i + \varepsilon_{ij1},$$

where the  $Z_{ij1}$  are first-period postnatal behavior of parents (for example, breastfeeding) and  $\varepsilon_{ij1}$  is the first-period stochastic component of health. Expression (2) embodies the assumption that the production of health is a cumulative process, with past inputs as well as past stochastic events having persistent effects.

In any dynamic optimizing model in which the health of children, and some subset of the inputs,  $Z_{ijk}$ , including family size, are arguments in the objective function, the demand for the prenatal inputs,  $Z_{ij0}$ , will be a function of prices  $p$ , income  $F$ , and  $\mu_i$ , namely,

$$(3) \quad Z_{ij0} = \psi(p, F, \mu_i).$$

The postnatal input levels  $Z_{ij1}$ , however, will also be functions of the realized stochastic health disturbances observed at the birth of the child; that is,

$$(4) \quad Z_{ij1} = \psi(p, F, \varepsilon_{ij0}, \mu_i).$$

In words, parents' consumption choices will reflect their awareness of the health en-

dowments of their children. Parents will also adjust their consumption behavior to perceived exogenous changes in any one of their children's health. For example, parents who expect to have children with a high risk of mortality (for example, low birth weight) might seek prenatal care earlier in their pregnancies and plan to have fewer (or more) children than parents who expect to have children with greater health endowments or who are to be born in less risky environments. Moreover, for given information on endowments, parents may also prolong breastfeeding or provide more resources to children who appear to be vulnerable or who contract an illness after birth.

The remedial and anticipatory behavior of parents means that, given that both  $\varepsilon_{ijk}$  and  $\mu_i$  are usually unobserved by the econometrician, the parental inputs that influence child health will not be uncorrelated with the residuals in the health production functions (1) and (2). That is,  $cov(Z_{ijk}, \mu_i) \neq 0$  for  $k = 0, 1$  and  $cov(Z_{ij1}, \varepsilon_{ij0}) \neq 0$ , although  $cov(Z_{ij0}, \varepsilon_{ij0}) = 0$ , since the random component of the birth outcome is by definition unforeseen by the parents during pregnancy. Ordinary least squares regression methods will thus not yield consistent estimates of the parameters of the health production functions, even if all inputs are observed. If consistent estimates could be obtained, however, the computed production function residuals would measure the health endowment with a random error. Regressing prenatal behaviors such as birth order on the production function residuals from either (1) or (2) would thus yield biased-to-zero estimates of the effects of exogenous changes in the anticipated family health endowment (genetic and environmental) of children. Estimates of the relationship between the health production function residuals and the behavior of the parents after the child's birth, however, would provide a mixture of the endowment effects and parental adjustments to new information acquired after the child's birth, since postnatal input behavior is a function of both  $\varepsilon_{ij0}$  and  $\mu_i$ , as in (4). Consequently, the pure expected endowment effects on health-related choices, such as of fertility, can only be assessed by analysis of prenatal behavior.

## II. Estimation Strategies

Consistent estimates of the production function for infant mortality provide measures of the biological effect of family size on mortality as well as the information needed to estimate the effect of an exogenous change in the health environment on parental fertility behavior. The theoretical framework suggests that such consistent estimates can be obtained by estimating the demand equations for the  $Z_{ijk}$ , (3) and (4), and by using the fitted values of the  $Z_{ijk}$  in estimating the mortality production function. Prices and income, as long as they are uncorrelated with the  $\mu_i$  and  $\varepsilon_{ijk}$ , serve as identifying variables, since such variables influence all health input and consumption choices, but do not directly affect mortality.

This two-stage estimation method has been applied to characterize the biological effects of parents' behavior on birthweight, gestation, and the rate of fetal growth (see our earlier paper). These estimates also show the significance of heterogeneity bias. However, prior work on the fertility-mortality relationship considering either heterogeneity or biological factors has not fully taken into account the stochastic-dynamic aspects of the health behavior by parents. Direct estimates of the effect of child mortality on fertility might be obtained using two-stage least squares, controlling for prices and income. However, such estimates (assuming that identification is theoretically justified), do not provide information on how parents would alter their fertility if the exogenous healthiness of their environment changes. Rather, the technique simulates an experiment in which parents are assigned a child death rate over which they have no control, that is, the effect of the mortality endowment cannot be altered by changes in household resource allocations, as it can be in the real world.

The biological determinants of child mortality, such as (1) and (2), have been estimated with explicit attention to parental or regional health heterogeneity by Randall Olsen and Kenneth Wolpin. In their study, a waiting-time regression method was used to estimate a child mortality production func-

tion based on within-family differences in children's time to death and in such postnatal health inputs as breastfeeding and child spacing. However, as can be seen from (2), if  $H_{i,j,1}$  is taken as a latent index of survivability, differencing across children only purges out the family/area endowment,  $\mu_i$ ; inputs will still be correlated with the residual in the fixed family-effects model, since the child effects,  $\varepsilon_{ij0}$ , will differ across children, are observed by parents after a child's birth, and will influence parents' postnatal behavior. As a consequence, Olsen and Wolpin must assume that parents do not adjust their input behavior in response to postnatal random shocks. The two-stage estimation approach which exploits variations in prices and income and the structure of the household demand model, however, yields production function estimates which are consistent in the presence of such dynamic behavior.

## III. Fertility and Infant Mortality: Estimates of Biological and Behavioral Relationships in the United States

In this section we discuss preliminary estimates of the biological effects of birth order on infant death in the United States and the effects of changes in the expected exogenous component of mortality on fertility and other types of parental behavior using the two-stage demand/production estimation procedure. The data are from the 1967, 1968, and 1969 National Natality Follow-back Surveys (USDHEW), which provide information on national probability samples of approximately 10,000 legitimate U.S. live births. The data indicate whether or not each child has died prior to the time when the parents responded to the survey questionnaire and the interval between the child's birth date and the date of the questionnaire (average of 19 months). The number of infant deaths reported is 209. From the survey information, six forms of behavior, four prenatal, were selected as potential biological determinants of infant mortality—birth order, delay after conception in seeking medical care during the pregnancy, mother's rate of smoking while pregnant, mother's age, duration of breastfeeding, and delay by the

mother after the child's birth before returning or going to work. In addition, the race of the mother and the child's sex were included as exogenous determinants of infant mortality.

To obtain the instrumental variables needed to estimate the mortality production function, state and county level information on health programs and prices were merged with the household socioeconomic and health data. The regressors in the demand equations for the six health-related behavioral variables included the schooling of the parents, the age-adjusted income of the father, local governmental health and hospital expenditures per capita, the number of hospitals and health departments with family planning services per capita, obstetrician-gynecologists per capita, prices of cigarettes and milk, the total and female unemployment rates, metropolitan area location and size, and the regional mix of employment by industry group (described more fully in our earlier paper). The set of variables explained a statistically significant proportion of the variance for all six inputs, with  $R^2$ s ranging from .03 for cigarette smoking to .14 for number of births and delay in obtaining prenatal medical care.

The two columns of Table 1 report the coefficients of the infant death production function estimated by ordinary least squares (*OLS*) and two-stage least squares (*TSLS*). Since the dependent variable is dichotomous and the residuals heteroscedastic, the reported *t*-values are not unbiased. Thus, while the *TSLS* parameter estimates should be consistent, statistical tests may be misleading.<sup>1</sup> Nevertheless, the *OLS* and *TSLS* coefficient estimates differ substantially. For example, while the *OLS* estimates suggest that delay in prenatal care has a small and *negative* effect on infant mortality, the *TSLS* results confirm the anticipated finding that such delay considerably increases the probability of infant death. More importantly, the inconsistent *OLS* parameter estimates indicate that higher birth order is associated with higher infant mortality while the older the

TABLE 1—ESTIMATES OF THE PRODUCTION FUNCTION FOR INFANT MORTALITY

Independent Variables (sample means)	Infant Mortality	
	<i>OLS</i>	<i>TSLS</i>
Birth Order <sup>a</sup>	.0129	-.0171
(2.51 children)	(10.4)	(2.46)
Prenatal Care Delay <sup>a</sup>	-.00145	.0332
(2.67 months)	(1.16)	(3.80)
Cigarettes Smoked <sup>a</sup>	.00002	-.00081
(4.68 cigarettes/day)	(.00)	(.42)
Mother's Age <sup>a</sup>	-.00119	.00304
(25.1 years)	(2.96)	(1.81)
Whether Breastfed <sup>a</sup>	-.00510	-.0316
(.260)	(1.28)	(1.51)
Whether Returned to Work <sup>a</sup> (.257)	-.00180	-.106
	(.45)	(2.75)
Black	.0290	.0474
(0.170)	(5.87)	(4.77)
Female	-.00742	-.00707
(.490)	(2.14)	(1.83)
Intercept	.0278	-.0615
	(2.02)	(3.32)
<i>F</i>	26.7	10.9
<i>N</i> = 8119		

Note: Absolute value of *t*-values in parentheses beneath regression coefficients. See text for their potential bias.

<sup>a</sup>Endogenous variable.

mother the less likely are the chances of an infant death. The *TSLS* estimates suggest just the opposite—higher fertility directly lowers child mortality and delay in child-bearing increases the mortality rate. Neither the *OLS* or *TSLS* estimates indicate that mother's smoking while pregnant is significantly related to the probability of child survival; this finding contrasts with other evidence that smoking reduces birthweight (see our earlier paper). The *OLS* estimates suggest there is no significant mortality relationship with mother's work after her birth, while *TSLS* estimates indicate that if the mother works, infant mortality is less, suggesting that benefits accruing from the mother's earnings outweigh any decrease in her time in child care. Black babies have considerably higher mortality rates, even when the black-white behavioral differences in mother's age at birth, fertility, smoking, breastfeeding and employment are "controlled."

While the heteroscedasticity of the residuals does not permit the conventional tests of the

<sup>1</sup>While logit estimates were calculated, the statistical properties of the two-stage logit estimates, where the first-stage equations are linear, are not well established.

statistical significance of heterogeneity bias (Durbin), the striking substantive differences between the *OLS* and *TSLS* estimates suggest that parents do respond to differences in their genetic/environmental health endowments. Further evidence is provided by separately regressing each of the prenatal inputs on the production function's residuals computed using the *TSLS* estimates and the actual inputs. The estimated coefficients were 1.57(13.3) for birth order,  $-1.89(20.2)$  for prenatal care delay, 1.75(3.2) for cigarettes smoked, and 1.23(3.5) for mother's age. (The *t*-statistics are given in parentheses.) These estimates, which provide lower bounds for the effects of the family-specific health endowment, are substantial.

In particular, the results suggest that the average number of children per mother would increase by one-sixth of a child if an infant mortality rate of 0.1 were anticipated. The positive associations between infant mortality and fertility seen in the gross correlations and in our *OLS* estimates appear to mask a *negative* biological effect of birth order on infant mortality ( $-0.0171$ ) and a substantial positive behavioral response of fertility to the anticipation of a higher mortality risk (1.57). If these response patterns are representative of historic levels, the decline in child mortality that has occurred in the last century in the developed countries could account for about one-fifth of the coincident decline in fertility.

The residual estimates also indicate that the responses to a *ceteris paribus* exogenous increase in anticipated infant mortality are not confined to fertility behavior; mothers appear to seek prenatal medical care significantly earlier when the exogenous rate is higher. In this instance the input response is compensatory, because it reduces somewhat the effect of the higher exogenous mortality.

Age at birth and mother's smoking, on the other hand, appear to decrease as the survival endowment of infants increases.

Finally, we might look for a biological-environmental basis for differences in black-white fertility and other input behavior, as our evidence suggested that black mothers experience exogenously higher infant mortality rates than do white mothers. However, despite the exogenously lower survival probabilities of black infants compared to white infants, black mothers obtain prenatal medical care *later* than do white mothers, even though our *TSLS* estimates indicate that such delay increases the risk of their child's death, and that in the total population, mothers with higher expectations of child mortality seek medical care earlier. Further study of this anomalous pattern of use of prenatal medical services by the black population is warranted.

## REFERENCES

- Durbin, J., "Errors in Variables," *Review of the International Statistical Institute*, 1954, 22, 23-32.
- Olsen, R. J. and Wolpin, K. I., "The Impact of Exogenous Child Mortality on Fertility: A Waiting Time Regression with Dynamic Regressors," *Econometrica*, March 1983, 51.
- Rosenzweig, M. R. and Schultz, T. P., "The Behavior of Mothers as Inputs to Child Health," in V. Fuchs, ed., *Economic Aspects of Health*, Chicago: Chicago University Press, 1982.
- U.S. Department of Health, Education and Welfare, National Center for Health Statistics, *Standardized Micro-Data Tape Transcripts*, DHEW Publication No. (PHS) 78-1213, June 1978.

# INTERNATIONAL DIMENSIONS OF MONETARY MANAGEMENT

## U.S. Monetary Policy and World Liquidity

By THOMAS D. WILLETT\*

Over the past decade, a new paradigm of international macroeconomics has emerged. Sometimes called global monetarism (see Marina Whitman), this approach differs from traditional monetarism by emphasizing open economy relationships in what is assumed to be a highly integrated world economy. In small open economies under fixed exchange rates, monetary expansion and contraction will lead primarily to balance of payments surpluses or deficits and ultimately to changes in national spending and prices. In such a world, quantity theory relationships break down at the national level while remaining valid at the global level. This paper focuses on the empirical importance of this paradigm.

### I. Global Monetarism and International Monetary Interdependence

While there are numerous precursors, the popularity of the global monetarist view is strongly associated with the reemergence of the emphasis on the monetary approach to the balance of payments and exchange rates determination led by Chicago economists such as Harry Johnson, Robert Mundell, and Arthur Laffer. The historical association between the huge U.S. balance of payments deficits and international liquidity explosion of the early 1970's and the accompanying substantial increase in the aggregate rate of monetary expansion in the major industrial countries was quite consistent with models of the reserve center's domination of world money supply determination under a Bretton Woods type of international monetary system. The international liquidity explosion of 1970-72 has been widely held to be the

major cause of the subsequent worldwide inflation.

With the breakdown of the Bretton Woods system, these initial global monetarist propositions lost their force. However, in simple monetarist models, flexible exchange rates insulate national economies from monetary disturbances abroad and convert monetary policy and inflation from international to national phenomena. Thus the strong monetary linkage between the United States and the rest of the world, typical of a fixed rate world, would be broken. While numerous studies written prior to the adoption of widespread floating in the 1970's had pointed to limitations on the extent to which flexible exchange rates could insulate national economies from one another, such limitations began to become much more widely appreciated after floating was widely adopted. Monetary policy changes often have real effects in the short run, and countries often care about the effects of exchange rate movements. Thus even under flexible exchange rates, significant monetary interdependence may remain. For example, even with complete control over the national money supply, both international capital flows and exchange rate movements can influence velocity through financial market and trade balance effects, and exchange rate movements can also affect short-run inflation-unemployment tradeoffs. Furthermore, the recent work on international currency substitution has stressed that flexible rates may not even allow countries to retain complete control over their national monetary aggregates.

Thus it comes as no surprise that Europe remains concerned about U.S. monetary and fiscal policy developments even under flexible exchange rates. The recent substantial appreciation of the dollar, caused in considerable part by U.S. monetary and fiscal

\*Claremont Graduate School and Claremont McKenna College.

policies, forced a dilemma on other countries. They must either tighten their own monetary policies or face a substantial depreciation of their currencies. Such a dilemma could not be avoided if a more stable economic environment were to be reestablished in the United States, although a different U.S. policy mix might have reduced the magnitude of this problem. There is a great deal of controversy about the quantitative strength of these short-run international linkages under flexible exchange rates. Recent econometric work has tended to yield a rather wide range of estimates, although most suggest lower levels of interdependence than are implied by popular political discussions. Still there can be little question that U.S. monetary policy has continued to have nontrivial short-run effects on the rest of the world even after the switch from fixed to flexible exchange rates.

A common feature of most analyses of international monetary relationships has been the assumption that impacts run almost exclusively from the United States to the rest of the world. In recent years this assumption has been increasingly challenged. While it is now widely recognized that international considerations cannot safely be ignored entirely in U.S. macroeconomic policymaking, several economists such as Laffer, Mundell, Ronald McKinnon, and Marc Miles have put forth the much stronger proposition that the United States should itself be viewed as a small economy which is dominated by international developments. Several of the recent arguments to this effect will be critically analyzed in Section III.

## II. The United States as the Determinant of Global Monetary Conditions

While initial global monetarist writings focused on the international liquidity explosion of 1970-72 as strong evidence for their hypothesis, more recent analysis has also focused on the heavy buildup of foreign official dollar holdings during the weakness of the dollar in 1977-78 as another example (see, for example, McKinnon, 1982). Again, at the aggregate level the story fits. For many countries a substantial increase in official

reserves was accompanied by a considerable increase in monetary expansion and subsequent acceleration of inflation. However, more detailed analysis weakens the strength of the argument. There is considerable truth to the posited linkage between international reserves and money supply changes as a long-run proposition required by the need to avoid persistent balance of payments disequilibrium over the long run. As a short-run proposition, however, this linkage rests on the inability or unwillingness of the national authorities to sterilize international reserve flows in order to keep them from influencing the domestic money supply.

In support of the global monetarist hypothesis, McKinnon cites estimates that sterilization is often less than complete. This is sufficient to support the proposition that the growth of international liquidity explosion will have a detectable influence on monetary expansion, but leaves open the crucial question of the magnitudes of these effects. Most of the recent estimates of sterilization coefficients for industrial countries suggest that they are typically well above .5 even for those countries that do not appear to have completely sterilized. Applying such estimates to the 1970-72 episode suggests that at most about one-third of the monetary expansion in the major industrial countries over this period could be attributed to the international liquidity explosion, with around 15 to 20 percent being a best guess. (See Leroy Laney's and my 1981 paper.) Direct estimates by Laney and myself suggest likewise that while the international liquidity effects were certainly not trivial in many countries, domestic influences were typically a good bit more important (see my 1980 study). Judgmental assessments in OECD and BIS reports also indicate the importance of domestic influences over this period.

While I have not yet completed analyzing the 1977-78 episode in detail, I strongly suspect that a similar interpretation will hold. There were strong domestic reasons for accelerated monetary expansion in many of the European countries and preliminary econometric work by Laney, Arthur Warga, and myself (forthcoming) finds that sterili-

zation coefficients have tended to rise further as countries have moved from pegged rates to managed flexibility.

### III. The United States as Denominated by World Monetary Conditions

One of the most common types of arguments that international influences have a dominant effect on U.S. monetary conditions focuses on the Eurodollar market. Periodically one sees popular articles which point out that the estimated size of the Eurodollar market is several times the size of  $M1$  in the United States and draw the conclusion that the Federal Reserve consequently can exert little control over U.S. monetary conditions. Such arguments overlook that the liquidity structure of the Eurocurrency market makes most of it credit, rather than money narrowly defined. Furthermore, the major conduit of international interbank lending does not substantially influence domestic monetary conditions until the funds are lent to domestic nonbanks. At this point they will show up in the domestic monetary statistics. Eurocurrency credit can of course also influence the velocity of the narrower aggregates, but strong evidence for the importance of such effects has not been presented (on these issues, see my 1980 study).

Eurocurrency transactions by nonbanks can give rise to domestic monetary influences, and some of these holdings are now included in the recent revisions of the definitions of U.S. monetary aggregates. While none of these transactions are judged to be comparable to the components of  $M1$ , overnight Eurodollar holdings are included with overnight repurchase agreements in  $M2$ . In 1980, however, these amounted to only about \$3 billion out of an  $M2$  total of over \$1,600 billion. The role of Eurocurrency holdings in the broad U.S. monetary aggregates has grown quite rapidly, however, from less than \$10 billion in 1975 to almost \$66 billion in 1982. These are included only in the broadest aggregate,  $L$ , which totaled over \$2,600 billion in 1981, although a case could be made on conceptual grounds for including them in  $M3$ , which totaled a little under \$2,200 billion. Eurocurrency transactions can

influence the money multiplier, but Anton Balbach and David Resler have estimated that this has "...only minor effects on the U.S. money stock" (p. 11). (The revision of U.S. monetary statistics in the work of the Bach Commission report included not only the addition of Eurodollar figures, but also deleted several categories of foreign holdings of demand deposits in U.S. banks on the grounds that these holdings typically did not seem closely related to economic and financial conditions in the U.S., see H. Farr et al.)

International transactions can also directly influence U.S. monetary conditions through effects on interest rates, currency substitution, and velocity. It has long been known that interest rates on comparable financial instruments in New York and the Eurocurrency markets move together, and that arbitrage opportunities are quickly eliminated. While it has generally been assumed that causation ran almost entirely from New York to Europe, in recent years this view has been challenged. It is certainly plausible to believe that with the growing relative size of the Eurodollar market (and of foreign holdings of U.S. government securities), international considerations can now have a nontrivial influence on U.S. interest rates. Relatively little work has been done so far, however, to estimate the magnitude of this influence. One application of Granger-Sims causality testing found greater causation running from Europe to New York than vice versa, but the applicability of the Granger-Sims methodology to this type of issue is open to considerable question. Furthermore, if the prevalent judgment of market participants is correct that genuine arbitrage opportunities are eliminated within minutes at most, then the lead-lag patterns revealed in Granger-Sims testing would apply basically to a statistical artifact due to the less than perfect compatibility of the data series.

In a similar vein, McKinnon (1981, 1982), has argued that exchange rate expectations have a dominant influence on U.S. interest rates. He illustrates his argument with the association of the substantial decline of the dollar in 1977-78 and the rise in U.S. interest rates over the same period. McKinnon makes no effort, however, to show that the

U.S. interest rate increase over this period cannot be adequately explained on domestic grounds and it is clear that at least some of the increase can be explained by the domestically generated rise in inflationary expectations over this period. (Of course, the fall of the dollar may have contributed further to the rise in inflationary expectations.) What is needed are attempts to investigate the role of international influences in domestic interest rate equations. I view this as an important area for research.

It has also been argued that international currency substitution has had a dominant impact on the dollar and U.S. monetary conditions, indeed to the point that even the United States is too small to be an independent currency area and hence should abandon flexible exchange rates (see Marc Miles). Laney, Christopher Radcliffe, and myself have argued that Miles' analysis rests on a failure to clearly distinguish between economic and statistical significance. We had no quarrel with Miles' finding of quite statistically significant currency substitution vis-à-vis the dollar, but noted that fluctuations in the data series he investigated were on the order of \$1 billion, a tiny fraction of U.S. *M1*. When his estimated elasticities of substitution are translated into the form of a standard demand for money function, the implied elasticities are quite small, on the order of .003.

Miles' study, however, investigated only one of the many possible channels for currency substitution and such substitution is certainly a possible explanation for the instability in U.S. demand for money functions which developed in the 1970's. The timing does not seem to match well, however, as McKinnon focuses on 1970-72 and 1977-78 and the associated weakness of the dollar as the major periods of currency substitution against the dollar, while most researchers have found that the domestic demand for money functions became unstable around 1974-75.

Bruce Brittain finds evidence of a significant negative correlation between movements of velocity around trend in the United States and Europe which would be consistent with major international shifts in currency

demands. Again, however, the whole story does not fit. The drop in U.S. velocity in 1977-78 and the rise in Europe is consistent with the posited currency substitution away from the dollar, but in the early 1970's, U.S. velocity for *M1* was well below trend. This would be consistent with currency substitution in favor of, rather than against, the dollar. McKinnon (1982) further argues that because of currency substitution, "In general, growth in the world money supply is a better predictor of American price inflation than is U.S. monetary growth" (p. 324). His supporting evidence is not convincing. He presents tables of annual U.S. and world money supply growth rates and inflation for inspection, but performs no formal statistical analysis. He appears to put considerable weight on the 1979-80 episode in which U.S. inflation was a good bit higher than would be expected on the basis of U.S. monetary expansion while the world money supply had been growing more rapidly.

Apart from the danger of extrapolating from one observation, it should be noted that one at least as equally convincing explanation of the high U.S. inflation rate was the substantial increase in oil prices over this period, McKinnon used wholesale price indices where oil prices are particularly heavily weighted. Furthermore, tests for currency substitution should focus on effects on the demand for money, velocity, or nominal spending, rather than just on prices since the latter can be confounded by shifts in inflation-output relationships. What is really needed for this type of investigation is the statistical comparison of various measures of U.S. and world money supplies in explaining nominal *GNP* or fluctuations in velocity in the United States. In our preliminary investigations, Radcliffe, Warga and I have not found strong evidence to support McKinnon's hypothesis. In general, McKinnon's world money supply series does not explain either the U.S. wholesale price index or nominal *GNP* better than U.S. *M1*, although it does do comparatively better for the *WPI* than *GNP*. For example, using current and two lagged values of percentage changes in U.S. *M1* or McKinnon's world money supply to explain the percentage change in the U.S.

*WPI* or nominal *GNP* and Hildreth-Lu corrections for serial correlation, we find that on the basis of  $\bar{R}^2$ , U.S. *M1* "out-explains" the world money supply by .44 to .35 for the *WPI* and by .73 to .60 for nominal *GNP*. The corresponding standard errors of estimate are .023 to .030 for the *WPI* and .013 to .016 for *GNP*. (Consistently we find the greatest explanatory power from money lagged one year.) Such regression results should certainly not be taken as definitive, but they should shed considerable doubt on the United States as a small country hypothesis.

#### IV. Conclusion

My conclusion is that the global monetarists have played a useful role of highlighting the potential importance of various aspects of international monetary interdependence. However, the currently available evidence does not support the strong propositions they have advanced about the dominance of U.S. monetary developments on the rest of the world or, conversely about the domination of U.S. monetary conditions by international developments. The actual strengths of these various types of monetary interdependencies should be important topics for further research.

#### REFERENCES

- Balbach, Anatol B. and Resler, David M., "Eurodollars and the U.S. Money Supply," *Federal Reserve Bank of St. Louis Review*, June/July 1980, 62, 2-12.
- Brittain, Bruce, "International Currency Substitution and the Apparent Instability of Velocity in Some Western European Economies and in the United States," *Journal of Money, Credit and Banking*, May 1981, 13, 135-55.
- Farr, H. et al., "Foreign Demand Deposits at Commercial Banks in the United States," in *Improving the Monetary Aggregates: Staff Papers*, Board of Governors of Federal Reserve System, November 1978.
- Laney, Leroy O. and Willett, Thomas D., "The International Liquidity Explosion and Worldwide Inflation: The Evidence from Sterilization Coefficient Estimates," *Journal of International Money and Finance*, No. 2, 1981, 1, 141-52.
- \_\_\_\_\_, Radcliffe, Christopher D. and Willett, Thomas D., "International Currency Substitution by Americans Is Not High: A Comment on Miles," *Claremont Working Papers*, 1982.
- \_\_\_\_\_, Warga, Arthur and Willett, Thomas D., "International Liquidity and Domestic Monetary Expansion," *Claremont Working Papers*, forthcoming.
- McKinnon, Ronald I., "The Exchange Rate and Macroeconomic Policy: Changing Postwar Perceptions," *Journal of Economic Literature*, June 1981, 19, 531-57.
- \_\_\_\_\_, "Currency Substitution and Instability in the World Dollar Market," *American Economic Review*, June 1982, 72, 320-33.
- Miles, Marc A., "Currency Substitution: Some Further Results and Conclusions," *Southern Economic Journal*, July 1981, 48, 78-86.
- Radcliffe, C. D., Warga, A. and Willett, T. D., "Global Monetarism and U.S. Monetary Conditions," *Claremont Working Papers*, forthcoming.
- Whitman, Marina, "Global Monetarism and the Monetary Approach to the Balance of Payments," *Brooking Papers on Economic Activity*, 3:1975, 491-536.
- Willett, Thomas D., *International Liquidity Issues*, Washington: American Enterprise Institute, 1980.

# Monetary Policy: Domestic Targets and International Constraints

By JACOB A. FRENKEL\*

Macroeconomic policies for open economies differ in fundamentally important ways from the corresponding policies for closed economies. The openness of the economy imposes constraints on the effectiveness and proper conduct of macroeconomic policies, and it also provides policymakers with information which may be usefully exploited in the design of policy. The discussion in this paper focuses on the dependence of monetary policy on the constraints and the information that are provided by the external sector. Section I summarizes briefly the characteristics of the international constraints on monetary policy. Section II deals with intervention in the foreign exchange market and its relation to monetary policy. In this context, the distinction between sterilized and nonsterilized interventions is drawn and the implications of the various forms of interventions for the effectiveness of monetary policy are examined. Finally, Section III addresses the question of the role that exchange rates should play in the design of monetary policy. It is argued that data from the market for foreign exchange in combination with data on interest rates can provide the monetary authorities with useful information on money market conditions and thereby can contribute to the improved conduct of monetary policy.

## I. The International Constraints

The open economy is linked to the rest of the world primarily through three key linkages: through international trade in goods and services; through international mobility

of capital; and through international exchanges of national monies (see my paper with Michael Mussa, 1981, for a detailed analysis of the implications of these linkages for macroeconomic policies).

International trade links prices in different national economies. While the evidence on purchasing power parities reveals that this link is not rigid, it is evident that a country cannot choose its long-run trend in the inflation rate independent of the long-run courses of monetary policy and the exchange rate. This relation thus imposes a severe constraint on monetary policy.

International mobility of capital links interest rates on financial assets. In addition, by permitting countries to finance current-account imbalances, it provides for a channel through which macroeconomic disturbances are transmitted internationally. The international mobility of capital limits the power of monetary policy. Under a fixed exchange rate regime, a monetary expansion in excess of money demand is likely to have only a limited success in sustaining the change in the nominal money stock. Any temporary reduction in the domestic rate of interest will induce capital outflow and a loss of foreign exchange reserves, and any attempts to sterilize the monetary consequences of the loss of international reserves is unlikely to be viable in the long run (more on this in Section II). Under a flexible exchange rate regime, the monetary authority regains control over the nominal money stock, but the international mobility of capital still imposes a severe limitation on the ability of monetary policy to significantly affect the evolution of output and employment. A monetary expansion is likely to induce a rapid change in the exchange rate which leads to prompt adjustment of prices and wages. The leverage of monetary policy can be somewhat enhanced if it operates in financial assets that are

\*University of Chicago and National Bureau of Economic Research. I am indebted to the National Science Foundation, grant SES-7814480 A01, for financial support. The research reported here is part of the NBER's Research Program in International Studies. Any opinions expressed are solely my own.

isolated from world capital markets since, in the short run, the link between the rates of return on such assets with the world rates of interest is not as tight.

The international exchange of national monies and the requirement of monetary equilibrium also impose a severe limitation on the effectiveness of monetary policy. As stated before, under a fixed exchange rate regime the authorities lose control over the nominal money stock, while under a flexible rate regime the requirement of monetary equilibrium ensures that in the long run, changes in the nominal money stock lead to a proportionate change in all nominal prices and wages. Because of the rapid change in the exchange rate, the constraint on monetary policy that is implied by the homogeneity postulate is likely to be manifested much more promptly in an open economy with flexible exchange rates than in a closed economy.

An additional consideration constraining the conduct of monetary policy follows from the dynamic linkage between current exchange rates and expectations of future exchange rates (see Mussa, 1976; 1979). This dynamic linkage implies that the effect of monetary policy on the exchange rate, and thereby on other economic variables, depends on its effect on expectations concerning future policies. These expectations, in turn, are influenced by the past and by the current course of policy, and it is likely that the mere recognition of this dynamic linkage will influence the conduct of policy. For the government, being aware that the effectiveness of any particular policy measure depends on the way by which it influences the public's perception of the implications of the measure for the future conduct of policy, is likely to be more constrained in employing the instrument of monetary policy.

In summary, the openness of the economy imposes constraints on monetary policy. These constraints are reflected in either a reduced ability to influence the *instruments* of monetary policy (like the nominal money supply under fixed exchange rates), or in a reduced ability to influence the *targets* of monetary policy (like the level of real output),

or in an increased prudence in the use of monetary policy because of the potentially undesirable effects on expectations.

The constraints on the conduct of monetary policy depend on the exchange rate regime. Therefore, the question of the country's *choice* of the optimal set of constraints on monetary policy can be answered in terms of the analysis of the choice of the optimal exchange rate regime. Such analysis reveals that the optimal exchange rate regime depends on the nature and the origin of shocks that affect the economy. Generally, the higher is the variance of real shocks which affect the supply of goods, the larger becomes the desirability of increased fixity of exchange rates. The rationale for the implication is that the balance of payments serves as a shock absorber which mitigates the effect of real shocks on consumption. The importance of this factor diminishes the larger is the degree of international capital mobility. On the other hand, the desirability of exchange rate flexibility increases the larger are the variances of shocks to excess supply of money, to foreign prices and to deviations from purchasing power parities (see my paper with Joshua Aizenman).

## II. Exchange-Market Intervention

The analysis of the international constraints on monetary policy is closely related to the analysis of the questions of whether the authorities can sterilize the monetary implications of the balance of payments and the monetary implications of interventions in the market for foreign exchange. Specifically, with respect to intervention, the difficulties in analyzing that question start with definitions since exchange-market intervention means different things to different people (see Henry Wallich). Some, especially in the United States, interpret foreign exchange intervention to mean *sterilized* intervention, that is, intervention which is not allowed to affect the monetary base and thus amounts to an exchange of domestic for foreign bonds. Others, especially in Europe, interpret foreign exchange intervention to mean non-sterilized intervention. Thus, for the Europe-

ans, an intervention alters the course of monetary policy, while for the Americans, it does not.

The distinction between the two concepts of intervention is fundamental and the exchange rate effects of the two forms of intervention may be very different depending on the relative degree of substitution among assets. In principle, sterilized intervention may affect the exchange rate by portfolio-balance effects (see Polly Allen and Peter Kenen, William Branson, and Dale Henderson), and by signaling to the public the government's intentions concerning future policies, thereby changing expectations (see Mussa, 1981). To the extent that sterilized intervention is effective in managing exchange rates, the constraint on the conduct of monetary policy would not be severe since the undesirable exchange rate effects of monetary policy could be offset by policies which alter appropriately the composition of assets. In practice, however, the evidence suggests that nonsterilized intervention which alters the monetary base has a strong effect on the exchange rate while an equivalent sterilized intervention has very little effect (see Maurice Obstfeld). These findings are relevant for both the theory of exchange rate determination and the practice of exchange rate and monetary policies. As to the theory, they shed doubts on the usefulness of the portfolio-balance model. As to the practice, they demonstrate that the distinction between the two forms of intervention is critical if the authorities mean to intervene effectively, and that it may be inappropriate to assume that the open economy constraints on monetary policy can be easily overcome by sterilization policies.

The preceding discussion defined interventions in terms of transactions involving specific pairs of assets. In evaluating these transactions it might be useful to explore the broader spectrum of possible policies. The various patterns of domestic and foreign monetary policies and foreign exchange interventions can be divided into three groups as follows: I: Domestic nonsterilized foreign exchange intervention ( $BM^*$ ); I\*: Foreign nonsterilized foreign exchange intervention ( $M^*B$ ); II: Domestic monetary policy ( $MB$ );

II\*: Foreign monetary policy ( $M^*B^*$ ); III: Domestic sterilized foreign exchange intervention ( $BB^*$ ); and III\*: Foreign sterilized foreign exchange intervention ( $B^*B$ ). This classification is based on the types of assets that are being exchanged. Thus, when the authorities exchange domestic money ( $M$ ) for domestic bonds ( $B$ ), the transaction is referred to as domestic monetary policy (as in II), while when the authorities exchange domestic bonds ( $B$ ) for foreign bonds ( $B^*$ ), the transaction is being referred to as domestic sterilized foreign exchange intervention (as in III).<sup>1</sup>

This general classification highlights two principles. First, it shows that the differences between the various policies depend on the different characteristics of the various assets that are being exchanged. These different characteristics are at the foundation of the portfolio-balance model. Second, it shows that domestic and foreign variables enter symmetrically into the picture. Thus, for example, a given exchange between  $M$  and  $B^*$  can be effected through the policies of the home country or through a combination of policies of the foreign country. This symmetry suggests that there is room (and possibly a role) for international coordination of exchange rate policies. It also illustrates the " $(n-1)$  problem" of the international monetary system: in a world of  $n$  currencies there are  $(n-1)$  exchange rates and only  $(n-1)$  monetary authorities need to intervene in order to attain a set of exchange rates. To ensure consistency the international monetary system needs to specify the allocation of the remaining degree of freedom (see Robert Mundell).

By and large the evidence on the effectiveness of sterilized intervention has been based on a comparison between patterns I and III within a single country framework. It is possible that some of the findings emerging from the single country studies may be modified once the foreign countries' behavior is taken into account. But, until presented with such

<sup>1</sup>A fourth possible policy would be the exchange of domestic money ( $M$ ) for foreign money ( $M^*$ ) rather than the exchange of domestic money for foreign bonds. Some have characterized this as pure foreign exchange intervention.

evidence, it is reasonable to conclude that it is very difficult to conduct effectively independent monetary and exchange rate policies.

### III. Exchange Rates and Monetary Policy

The recent volatility of exchange rates and the large divergence from purchasing power parities have given rise to various proposals concerning rules for intervention in the foreign exchange market. Some of these proposals are variants of a purchasing power parity (*PPP*) rule according to which the authorities are expected to intervene so as to ensure that the path of the exchange rate conforms to the path of the relative price levels. In view of the discussion in Section II, these proposals, if effective, amount to guidelines for the conduct of monetary policy.

There are at least four difficulties with a *PPP* rule. First, there are intrinsic differences between the characteristics of exchange rates and the prices of national outputs. These differences, which result from the much stronger dependence of exchange rates (and other asset prices) on expectations, suggest that the fact that exchange rates have moved more than the price level is not sufficient evidence that exchange rate volatility has been excessive.

Second, the prices of national outputs do not adjust fully to shocks in the short run, and thus, intervention in the foreign exchange market to ensure purchasing power parity would be a mistake. When commodity prices are slow to adjust to current and expected economic conditions, it may be desirable to allow for "excessive" adjustment in some other prices.

Third, there are continuous changes in real economic conditions that require adjustment in the equilibrium relative prices of different national outputs. Under these circumstances what seem to be divergences from purchasing power parities may really reflect equilibrating changes.

Fourth, if there is short-run stickiness of prices of domestic goods in terms of national monies, then rapid exchange rate adjustments, which are capable of changing the

relative prices of different national outputs, are a desirable response to changing real economic conditions. An intervention rule which links changes in exchange rates rigidly to changes in domestic and foreign prices in accord with purchasing power parity ignores the occasional need for equilibrating changes in relative prices.

While it might be tempting to "solve" the problem of divergences from *PPP* by adopting a rigid *PPP* rule, this would be a mistaken policy course.

What should be the role of the exchange rate in the design of monetary policy? Generally, given that monetary and exchange rate policies should not be viewed as two independent instruments, consideration of the external value of the currency should play a relatively minor role in the design of monetary policy. The major consideration that should guide the monetary authority is that of achieving price stability.

While this prescription may seem to represent a revival of the "benign neglect" attitude the opposite is the case. In the past, one of the major arguments for the benign neglect attitude in the United States was that the U.S. economy was relatively closed and the foreign trade sector was relatively unimportant. The typical statistic which was used to justify this position was the low share of imports in *GNP*. This argument was inappropriate in the past and is even less appropriate under present circumstances. The United States has always been an open economy. The relevant measure of openness to international trade in goods and services is not the share of actual trade in *GNP* but rather the share of tradeable commodities in *GNP* (i.e., of potential trade) which is by far larger than that of actual trade. Furthermore, as stated in Section I, one of the main linkages of the United States to the world economy is operating through world capital markets with which the United States is clearly well integrated. The same principle applies to the measures of openness of most countries.

The prescription is based on the notions that the economy *is* open, that the external value of the currency *is* important, that the restoration of price stability is an important

policy goal, and that policy which views the exchange rate as an independent target or, even worse, as an independent instrument, is likely to result in unstable prices. Furthermore, if monetary policy succeeds in achieving price stability, it might be useful to allow for fluctuations of the exchange rate which provide for a partial insulation from misguided foreign monetary policies.

Even when monetary policy is not guided by exchange rate targets, it might attempt to offset disturbances arising from shifts in the demand for money. Such shifts in demand may be especially pronounced under a regime of flexible exchange rates. A policy which accommodates such demand shifts by offsetting supply shifts would reduce the need for costly adjustments of exchange rates and national price levels. The difficulty with implementing this policy is in identifying when a shift in money demand has occurred. As is obvious, the nominal rate of interest is not a reliable indicator of money market conditions. The more relevant indicators are the components of the nominal rate of interest—the real rate of interest and the expected rate of inflation—but these components are unobservable.

Here the exchange rate may be useful as an indicator for monetary policy especially when frequent changes in inflationary expectations make nominal interest rates an unreliable indicator of fluctuations in money demand. Accordingly, a *combination* of a high nominal interest rate differential and a depreciation of the currency, that seems to have prevailed in the United States during most of the 1970's, may have indicated a rise in inflationary expectations, which should obviously not have been fueled by an increase in the supply of money. On the other hand, a *combination* of a high nominal interest rate differential and an *appreciation* of the currency, that seems to have prevailed since the latter part of 1979, may indicate a rise in the demand for money, which *should* be accommodated by an expansionary monetary policy (this argument draws on my papers with Mussa, 1980, 1981, and my 1981 paper).

This prescription that is based on the relation between exchange rates and interest rates

can also shed light on the recent controversy concerning the proper conduct of U.S. monetary policy in view of the high rates of interest that have prevailed since 1980. The relatively tight monetary policy which accompanied the high nominal rate of interest in the United States was justified on the grounds that the high nominal rate of interest was primarily due to high inflationary expectations. As a counterargument, it was argued that the prime reason for the high nominal rate of interest was the high real rate rather than inflationary expectations. Obviously, the two alternative prescriptions call for fundamentally different monetary policies. To combat inflationary expectations, monetary policy had to be tight, but to combat high real rates of interest, a case could be made for a more relaxed monetary policy.

Here again the relation between the exchange rate and the rate of interest can provide the monetary authority with information that can be helpful in solving the "signal extraction" problem. By and large, since the latter part of 1979, the high nominal rate of interest in the United States has been accompanied by an appreciation of the dollar. This suggests that the important factor underlying the evolution of the nominal rate of interest in the United States has been the evolution of the real rate of interest rather than inflationary expectations. Under such circumstances the U.S. monetary policy could have afforded to be more relaxed while paying even more attention to the underlying reasons for the high real interest rates. Several factors have contributed to the rise in real interest rates. First, there have been large current and prospective budget deficits in the United States and in the rest of the world.

Second, stagflation lowered the hedging quality of bonds. With a weak economy and high inflation, the real interest rate on bonds declines. For bonds to be more attractive to bondholders, they must bear a higher real yield.

Third, high real interest rates represent a rise in the risk premium, attributable to several factors: (a) the projected rise in future budget deficits creates uncertainty about how these deficits will be financed; (b) the

volatility of monetary policy since late 1979 may have induced a rise in the risk premium; and (c) the fragility of the world financial system, the sequence of banking crises, the increased perception of sovereign risk and increased sensitivity to large exposures, and the increased reluctance to extend additional credit have all contributed to the rise in the risk premium and in real interest rates. This rise in risk has been reflected in the increased spread between high- and low-quality bonds.

Fourth, it has been argued that changes in the laws dealing with the treatment of depreciation and in those dealing with bankruptcies have also contributed to the rise in real interest rates.

This perspective suggests that monetary policy can use the information provided by the foreign exchange market to identify the source of variations in nominal rates of interest. Thus, the external sector while imposing severe constraints on monetary policy, is also providing it with useful information.

## REFERENCES

- Allen, Polly R. and Kenen, Peter B., *Asset Markets, Exchange Rates and Economic Integration*, Cambridge: Cambridge University Press, 1980.
- Branson, William H., "Assets Markets and Relative Prices in Exchange Rate Determination," *Sozialwissenschaftliche Annalen*, No. 1, 1979, 69–89.
- Frenkel, Jacob A., "Flexible Exchange Rates, Prices, and the Role of 'News': Lessons from the 1970's," *Journal of Political Economy*, August 1981, 89, 665–705.
- \_\_\_\_\_ and Aizenman, Joshua, "Aspects of the Optimal Management of Exchange Rates," *Journal of International Economics*, November 1982, 13, 231–56.
- \_\_\_\_\_ and Mussa Michael L., "The Efficiency of Foreign Exchange Markets and Measures of Turbulence," *American Economic Review Proceedings*, May 1980, 70, 374–81.
- \_\_\_\_\_ and \_\_\_\_\_, "Monetary and Fiscal Policies in an Open Economy," *American Economy Review Proceedings*, May 1981, 71, 253–58.
- Henderson, Dale, "Modeling the Interdependence of National Money and Capital Markets," *American Economic Review Proceedings*, May 1977, 67, 190–99.
- Mundell, Robert A., *International Economics*, New York: Macmillan, 1968.
- Mussa, Michael L., "The Exchange Rate, the Balance of Payments and Monetary and Fiscal Policy Under a Regime of Controlled Financing," *Scandinavian Journal of Economics*, May 1976; reprinted in Jacob A. Frenkel and Harry G. Johnson, eds, *The Economics of Exchange Rates: Selected Studies*, Reading: Addison-Wesley, 1978.
- \_\_\_\_\_, "Empirical Regularities in the Behavior of Exchange Rates and Theories of the Foreign Exchange Market," *Policies for Employment, Prices, and Exchange Rates*, Vol. 11, Carnegie-Rochester Conference Series on Public Policy, *Journal of Monetary Economy*, Suppl. 1979, 9–57.
- \_\_\_\_\_, "The Role of Official Intervention," Occasional Paper No. 6, New York: Group of Thirty, 1981.
- Obstfeld, Maurice, "Exchange Rates, Inflation and the Sterilization Problem: Germany, 1975–1981," *European Economic Review*, forthcoming 1983.
- Wallich, Henry C., "Exchange-Market Intervention: Issues and Views," *Journal of Commerce*, August 12–13, 1982.

# Internationally Managed Moneys

By GEORGE M. VON FURSTENBERG\*

This paper attempts to clarify the national policy content of international approaches to money supply management. It starts with a taxonomic section and then discusses three of the possible forms of international management of national money supplies. Elements of such management can be achieved by committee, by rules, or by inventing a stateless money which has a chance of becoming dominant over national moneys in international use. Prospects for attempts to bring greater international influence to bear on national money supply processes are appraised in the concluding section.

## I

Robert Triffin (pp. 146–47) has described the surrender of national sovereignty that would be implied by moving toward an internationally managed system as involving a trade of national policy instruments against international or supranational policy instruments adequate to serve the broad objectives of economic policy in the modern world. By this standard, there are no full-fledged internationally managed moneys. For instance, in the framework of this paper, the SDR and the ECU, which are created by international agreements on allocations or swaps, would not be viewed as such moneys on account of their restricted use in official transactions, their scant leverage over national policies, and their limited substitutability for foreign exchange. In fact, internationally managed moneys may never have existed in the full sense in which this term has been understood. For instance, for a long time, gold was a supranational money, but its supply and the resulting supply of bank money were

never managed internationally (Richard Cooper; John Maynard Keynes discussed methods that could be used to achieve supranational management in this regard). Under the gold exchange and dollar standards, the dollar, while at first freely convertible into gold at a fixed official price, continued to be managed nationally in such a way that increased monetary independence from the policies of the United States eventually became one of the siren calls for abandoning fixed rates. Since then the disintegration of central control structures has been accompanied by growing internationalization of and through markets (Jacques Polak), but the question of individual vs. joint pursuit of policy objectives and instrument settings remains.

Having winnowed the concept of internationally managed moneys, what is the grist that could remain in each of the three categories mentioned at the outset?

1) In theory, money could be managed by an international committee. If this is to be done in much the same way in which the U.S. money supply is managed by the collective of governors of the different Federal Reserve districts, a monetary union would first have to be established for that purpose (for a framework to evaluate the costs and benefits of participation, see Koichi Hamada). If it is done by negotiating money growth among governments represented in an international institution, the authority and decision rules of that institution would determine whether national money supply growth is managed internationally in substance.

The money supply decisions of one country will tend to influence those of other countries under almost any system that may be applied in the modern world. If the ways in which these repercussions feed forward through the reaction functions of other countries cannot be controlled or modified by them in important respects, for instance, be-

\*Chief of the Financial Studies Division, Research Department, International Monetary Fund. I am indebted to Benjamin J. Cohen, Richard N. Cooper, Robert E. Cumby and Herbert G. Grubel for helpful suggestions and comments. The usual disclaimer of institutional responsibility applies.

cause of binding and "unalterable" commitments to fixed exchange rates with the country initiating monetary change, and there are no significant feedbacks that could create mutual dependency, the result is hegemonial money. If there are no such commitments, discretionary management of national money supplies would have to involve more than situational attempts to coordinate policies and to resolve crises if it is to be international in character.

2) International money management could be conducted by rules rather than by discretion. If these rules are to govern national money supply growth they must be tied to criteria of economic performance. A system of fixed exchange rates does not, by itself, prevent countries from pursuing various inflationary policies jointly, even if they cannot choose to do so on their own without jeopardizing exchange rate fixity. Hence, the monetary antidotes to any price level and exchange rate movements that may occur would have to be agreed to in advance and be implemented predictably if independent targeting were to be superseded by the adoption of a suitable international convention (Ronald McKinnon).

More modestly, there may be particular facilities or mechanics of operation which are supported by internationally agreed rules that create rewards for certain kinds of money supply management and penalize others. Incentives that are embodied in concrete facilities and accepted instruments are not as easily disestablished or denied as ambitious global rules. For instance, if the governments of major countries all agreed to issue purchasing power bills, or bonds indexed to their respective price levels, they might become less inclined to generate inflationary surprises, and their residents might feel less obliged to diversify their portfolios internationally to hedge against such surprises.

3) Finally, internationally managed moneys could conceivably arise not from discretionary agreements or firm conventions negotiated among governments, but by national and international sponsorship—or, at least, legal toleration—of a stateless money, such as a currency-option money. (Opportunities for sponsoring a commodity money

are not considered in this paper.) If such an alternative money should be able to compete successfully with national moneys in private use, it might help establish international monetary order by disciplining the growth rate of national moneys.

As development of Euromarkets has suggested, financial innovations that are neither centrally planned nor even officially encouraged can create new facts for the international monetary system and for the way in which money supply impulses in one country are transmitted to and processed by others. The question for stateless moneys is whether they could provide disciplines similarly to those (vainly) sought under the gold standard by meeting the demand for a unit of stable universal purchasing power better than all or most national moneys.

## II. Discretionary Management

Major countries are aware that there may be substantial costs of pursuing policies whose implications for inflation differ sharply from the inflation rates expected in other major countries. These costs include increased volatility of the terms of trade and increased risks of political and macroeconomic instability, government interference, and distortions. Although neither the costs nor the hoped-for domestic benefits of divergence are easy to quantify or to add up over time, most governments attach some value to international compatibility of policies. They thus seek to influence the policies of other countries as these countries attempt to influence theirs. Put differently, the set of choices that are feasible for each country is likely to contain some elements that are particularly injurious for other countries. Countries can therefore benefit by offering to exclude those choices in return for other countries' accepting equivalent restraints on their freedom of action in different areas.

Pressures resulting from the experience of interdependence tend to be conveyed with the help of positive images, such as the locomotive or convoy similes of the 1970's, and by appealing to international comity. Appeals are based more on notional deviations from economic principles and objec-

tives to which most major countries appear to have subscribed in a general way than on demonstrable violations of codified rules. Mutual surveillance and international advice help crystallize and apply whatever standards appear to be acceptable to a qualified majority of countries as a basis for collective appraisal. Whether that appraisal is powerful in causing national policies to be modified depends in part on the degree to which costs of ignoring international advice can be brought home to the country concerned, and on the strength and cohesion of the underlying coalition on which this advice rests. The more countries differ on the basic principles of analysis that ought to be applied, on the ranking of policy objectives, and on the time frame relevant for their achievement, the less can negotiations be expected to contribute to international management of money.

### III. Application of Rules

Countries may be willing to agree to narrow the range of policy choices available to them, if the freedom of action in other countries is curtailed likewise. If the rewards to be sought or damages to be avoided can affect many countries reciprocally, adopting a convention of rules could generate greater predictability of future policies and of policy responses to unfolding events, thereby reducing the potential dispersion of standards and policies among countries. In areas where cooperation can create new goods, countries may also be able to enhance their opportunities and security by reaching agreements that are self-enforcing because the loss of benefits from ceasing to cooperate would be plain to each.

In the trade area, a rules approach would be exemplified by GATT. The efficacy of that agreement is protected by countries being allowed to retaliate unilaterally against violators and to deny most-favored-nation treatment to countries which do not subscribe to the rules of GATT. Other examples of common interests whose pursuit can be organized through the adoption of rules based on reciprocity relate to the sharing of energy supplies in emergencies, opening government procurement to foreign bidders

and limiting subsidies on international trade credit.

It is much more difficult to devise rules in the monetary area that can be protected by credible threats of retaliation and discrimination against violators. Since there may be no effective way for countries to pressure others engaged in excessive money creation—though exchange markets may react by depreciating the currency involved by much more than in proportion to its extra rate of money creation, thereby inaugurating a vicious cycle—international rules may not be enforceable. In that case, any shared commitment to adhere to money supply growth compatible with domestic price stability or related objectives would be predicated on unanimity, that is, voluntary compliance.

To increase international leverage, a system of mutual dependency can conceivably be constructed from the ground up by specific undertakings of a single country. Thus, the United States could pledge itself to rules which allow other countries to bring pressure on the United States to adjust its policies, thereby allowing the rules adopted by the country to be complemented by those of others. The end result could be an internationally managed, or self-managing, system that “neither requires the creation of new international institutions nor the assumption of new functions by existing ones” (Friedrich Lutz, pp. 16–17). For instance, when the United States stood ready to exchange dollars for a fixed quantity of gold and other countries endeavored to maintain fixed exchange rates with the dollar, they could, up to a point, attempt to discourage inflationary policies in the United States by threatening to convert dollars into gold (for details on how this threat has been appraised in past literature, see Robert Cumby). In due course, some countries other than the United States might then even have offered convertibility as Lutz envisioned, thereby generalizing the gold exchange standard in a multiple currency system.

### IV. An Alternative to National Moneys

In theoretical discussions, money has commonly been described as a commodity or an

asset with a comparative advantage in absorbing and disseminating information concerning transactions (see, for instance, Karl Brunner and Allan Meltzer). Hence stability, predictability, and uniformity of purchasing power in the international market would be among the prerequisites for a national money to emerge dominant. A dominant money, in turn, exerts pressure on other countries to compete, for instance by pegging the exchange value of their currency to the dominant money or by trying to achieve even lower rates of domestic inflation than the country which issues it.

To stimulate such competition without relying on a single national money and the policies pursued with it, international sponsorship of a currency-option money could be contemplated by which such a unit could be enabled to become a legally accepted denominator for a wide variety of private claims. Such a money would retain its purchasing power to a greater degree than all or most national moneys on account of the option feature whose pricing has recently been discussed by René Stulz. A unit of currency-option money is here defined as being worth no less than  $\bar{x}_i$  in any of  $i$  currencies, where the value of the fixed amounts  $\bar{x}_i$  could be equal at the point of introduction at the exchange rates then prevailing. Subsequently, however, when exchange rates have changed, the redemption value of the option money in the  $j$  currencies other than the one (or more than one, in a joint float) that has appreciated most will be higher. Once  $x_j$  has moved far above the contractual minimum  $\bar{x}_j$ , a minimum of this type could be raised at periodic intervals. This could be done without loss of continuity in redemption value to restore the insurance value of the currency option which was lost when the market value of the currency-option money fell towards its exercise value.

Because of its versatility and unexcelled maintenance of purchasing power, currency-option money could become an attractive denomination in international hedging and investment applications. If national authorities were to allow its use to spread, they might be induced to compete for the status of their money contributing the effective  $\bar{x}_i$

at which the redemption value of the currency-option money remains fixed. Such competition could reduce inflationary biases in the policies of major countries. Should the U.S. dollar consistently appreciate most against other major currencies in the future, currency-option money might become interchangeable with dollars at a fixed rate. Nevertheless, if international transactions and claims continued to be denominated in currency-option money in significant volume, the U.S. dollar standard that might be reaffirmed de facto would be put on indefinite probation.

### V. Prospects

The preceding discussion has suggested that elements of international management of national money supplies could be strengthened if means were found to make inflationary policies more costly for major countries. However, if a compelling majority of such countries cannot be found which desire to emulate the price performance of the least-inflating country among them, the degree of international influence that can be brought to bear is likely to be small. Under such circumstances no country would accept rules, international arrangements, or instruments raising the penalties for poor performance unless it could be convinced that there is an adequate *quid quo pro* other than putting it on its good behavior. Little progress would then be expected in any of the directions that could lead to greater international control of money for and through the group of countries whose currencies figure prominently in international finance.

Nevertheless, changes conducive to such a development can be identified. For instance, there would be greater opportunity for internationally coordinated management if those few countries who pay more than lip service to the goal of first attaining and then maintaining approximate stability of their domestic price levels would bring their combined weight to bear on advancing that cause regardless of any regional groupings to which they may belong. The formation of international issue coalitions intended to promote joint pursuit of a single major objective

should not be impeded by fixed regional groupings which limit the freedom of coalition for some of their members.

Even if, under the system of floating, inflation had its roots in national monetary, fiscal and exchange rate policies, and not in the international monetary system per se (Gottfried Haberler, p. 131), the effects of all or most countries' stabilization efforts have become increasingly predicated on the policies pursued elsewhere (McKinnon). On account of this partial erosion of independent control, cooperation in the use of monetary policy instruments may become attractive to make the achievement of national objectives more certain and less costly for each of the countries sharing nonconflicting goals, such as a return to approximate stability of their domestic price levels. At the same time, countries should be convinced that they cannot expect cooperation or foreign support for their currencies if they should slip into inflationary experiments. Refusal to accommodate any such shocks internationally would be vital to the defense of a noninflationary coalition. Nevertheless, even under the best of circumstances it remains unsettled whether an effective issue coalition could be maintained for very long if the United States should cease, once again, to be its leading member.

#### REFERENCES

- Brunner, Karl and Meltzer, Allan H., "The Uses of Money: Money in the Theory of an Exchange Economy," *American Economic Review*, December 1971, 61, 784-805.
- Cooper, Richard N., "The Gold Standard: Historical Facts and Future Prospects," *Brookings Papers on Economic Activity*, 1:1982, 1-45.
- Cumby, Robert E., "Special Drawing Rights and Plans for Reform of the International Monetary System: A Survey," in *International Money, Credit and the SDR*, Washington: International Monetary Fund, forthcoming 1983.
- Haberler, Gottfried, "How Important is Control over International Reserves?," in R. A. Mundell and J. J. Polak, eds., *The New International Monetary System*, Washington: International Monetary Fund and Columbia University Press, 1977, 111-32.
- Hamada, Koichi, "On the Political Economy of Monetary Integration: A Public Economics Approach," in R. Z. Aliber, ed., *The Political Economy of Monetary Reform*, New York: Landmark Studies, 1977, 13-31.
- Keynes, J. M., "Problems of Supernational Management," *A Treatise on Money*, ch. 38, 1930; reprinted in *The Collected Writings of John Maynard Keynes*, Volume VI, London: Macmillan and St. Martin's Press, 1971, 348-67.
- Lutz, Friedrich A., *The Problem of International Liquidity and the Multiple-Currency Standard*, Princeton Essays in International Finance, No. 41, International Finance Section, Princeton University, March 1963.
- McKinnon, Ronald I., "U.S. Monetary Policy and the Stability of the World Economy," unpublished paper, Stanford University, September 1982.
- Polak, Jacques J., "Coordination of National Economic Policies," Occasional Paper 7, New York: Group of Thirty, 1981.
- Stulz, René M., "Options on the Minimum or the Maximum of Two Risky Assets: Analysis and Applications," *Journal of Financial Economics*, July 1982, 10, 161-85.
- Triffin, Robert, *Gold and the Dollar Crisis*, New Haven: Yale University Press, 1961.

## EXPLORING BLACK WELFARE DEPENDENCY

### Changes in Black Family Structure: Implications for Welfare Dependency

By WILLIAM DARITY, JR. AND SAMUEL L. MYERS, JR.\*

Female headship among black families long has been more pronounced in the United States in comparison with other ethnic groups. E. Franklin Frazier's classic study of the black family in the 1930's placed a distinct emphasis on the disproportionately high number of "urban Negro families with women heads." Frazier's work suggested that throughout the pre-World War II period almost one-quarter of black families were headed by women.

In the mid-1960's, female headship among black families was the subject of Johnson Administration policy planner Daniel Moynihan's notorious characterization of the black family as enmeshed in a "tangle of pathology." But while the subsequent debate between disciples of Moynihan's "pathology-disorganization perspective" and the proponents of the "strength-resiliency perspective" raged, the proportion of black families with female heads has risen markedly. The proportion climbed from slightly less than 25 percent in 1965 to an astonishing more than 40 percent by 1980. Female headship also has grown among white families, but the rate of increase has not approached that among blacks. Between 1965 and 1980, the percent of white female-headed families rose from 9 percent to close to 12 percent.

Female-headed families, regardless of race, are uniformly poor in economic resources. In

1978, for example, while 36 percent of white families that were female-headed with related children under 18 years had incomes below the poverty line, the figure was 59 percent for black families. Female-headed families, especially *black* female-headed families, are viewed increasingly as a major element of a "permanent underclass."

The poverty of families with women heads has directed attention to the connection between family structure and the reliance of a significant part of the U.S. population on the welfare system, particularly the Aid to Families with Dependent Children (AFDC) program. Heather Ross and Isabel Sawhill, for example, have devoted an entire chapter of their book on female-headed families to the relationship between welfare and female headship. Two recent studies, one undertaken by Martin Rein and Lee Rainwater and a second by Richard Coe, have used longitudinal data from the University of Michigan's Panel Study of Income Dynamics to explore statistically the extent of dependence on welfare in the United States. Both studies concluded that welfare "dependence" has been exaggerated—although there is far more ambiguity to the results reported in each study than the authors suggest. Regardless, Coe (p. 46) reports that the long-term welfare population, those on welfare eight to ten years, is disproportionately composed of blacks, particularly nonelderly black females with young children. Ross and Sawhill also noted the "longer duration of welfare reciprocity for blacks" (p. 114). Welfare dependency thus is viewed more and more as a problem far more prevalent among black families.

#### I. Black Dependency Viewed Historically

The Ross-Sawhill, Rein-Rainwater, and Coe studies provide operational definitions

\*University of Texas at Austin, and University of Pittsburgh and Federal Trade Commission, respectively. Our research was supported in part by funds made available through the Center for Economic Policy and Welfare Reform at Central State University under grant no. 73-39-81-06 from the Employment and Training Administration, U.S. Department of Labor, under the authority of Title III, part B of the Comprehensive Employment and Training Act of 1973, as amended. Points of view or opinions stated in this document do not necessarily represent the official position or policy of the Department of Labor.

of dependency specific to the recent expansion of social transfer programs. Their studies see the current instance without historical perspective. Therefore, there is no analysis offered of the consistent client status imposed on black Americans from the nation's beginnings.

During slavery times, the southern plantocracy consciously contrived to keep blacks from developing a self-reliant basis for group support. The range and variety of institutions developed by either slaves or freed blacks were supervised carefully to preclude genuine ethnic autonomy. Black access to property and education also fell under the control of the planter elite, as did the conditions shaping the development of black family life. Both the fragmentation and the stabilization of family units was contingent in part upon the perceived requirements of the planters at different stages of the evolution of the slave system (see especially Frazier, Part One).

On the eve of the Civil War and "emancipation," black dependency underwent a further transformation. From a state of extreme clientage vis-à-vis the plantocracy, blacks increasingly became wards of the national judiciary. The most visible expression of this trend was the well-known *Dred Scott v. Sandford* decision handed down by the U.S. Supreme Court in 1857 which communicated the principle that the black man had no rights that the white man *must* respect. In its denial of citizenship to Dred Scott and his family, the decision rendered the legal protection available to black families generally subject to the arbitrary and transitory whims of the courts. Loren Miller, in a profound study of the peculiar relationship between blacks and the judiciary, published in 1966, observed the upshot of the *Dred Scott* decision was the following:

... the Court... had made Negroes, free and slave, its ward by attributing to the Constitution a complete scheme defining and regulating their relationship with the federal government, the state, and the body politic, and by arrogating to itself the exclusive power to assess

the nature, character, and extent of these relationships. [p. 79]

Miller also argued "... the guardian-and-ward relationship between the Supreme Court and the Negro, originating before the Civil War, has persisted into our times" (p. 13). The Court in the immediate post-Civil War period invoked the primacy of "states rights" where "the Negro" was concerned to sterilize the intent of the Thirteenth, Fourteenth, and Fifteenth Amendments (see Miller, pp. 102-17). "Civil rights" litigation did not appear to take a pro-black thrust until the late 1950's under the Warren court. But even if such change was more real than apparent, it did not alter the fate of blacks as reliant, in the last instance, on the preferences of the nine judges presiding on the Court at that particular moment.

Moreover, the advance of civil rights protections in the 1960's for individuals occurred while the Court made decisions that undermined the capacity of women as single parents—especially black women—to support their families. In a brief review of the development of the law relevant to the economic resources available to women, with children, divorced or separated from their spouses, Ross and Sawhill (pp. 94-97) indicate that the courts have made AFDC the primary source of support if the woman is not a wage earner. The courts largely have sustained a diminution of the obligations of the absent spouse. Further revealing the contradictory pattern of the Supreme Court's actions in the 1960's, the Court refused to review a lower court decision in 1966 that effectively "struck down congressional legislation passed in 1962 designed to protect women and children from loss of income and property associated with divorce or desertion" (Darity, p. 21).

Inexorably, black female-headed families have been pushed toward the welfare system—toward welfare *dependency* as the most recent manifestation of historic black dependency. Simultaneously concern has grown over the welfare "crisis" and the need articulated for welfare "reform." This reveals in acute form a diminished perception that the black dependent population is socially

necessary. Black dependency under slavery was desired by the larger society; black dependency today is increasingly seen as a burden by the larger society.

## II. Economic Theory, Black Families, and Welfare Dependency

Lacking the historical perspective advanced in the previous section, economists have treated the relationship between black family structure and welfare dependency solely within the rubric of the so-called "new economics of the family." The growth in female-headed families generally is attributed to mutually rational choices made by men and women. Some variants of the economic model place the onus of decision solely on the women (see, for example, Sheldon Danziger et al.). Their personal constrained optimization calculus leads more and more women to choose a preferred income-leisure package associated with single parenthood.

The major progenitor of the new economics of the family, Gary Becker, has reduced the entire process of family formation and dissolution to a branch of applied microeconomics. He has argued that female headship among blacks has grown due to black women perceiving economic advantages from AFDC income compared with the incomes that they or black men could earn in the labor market.

Under the intellectual influence of Becker and Jacob Mincer, Majorie Honig undertook a major empirical study in the early 1970's to examine the implications of the microeconomics of the family. She observed in time-series data for the decade 1960-70 increasing female headship while AFDC payments rose relative to the mean wage earned by males in manufacturing. She postulated a connection, but, oddly enough, tested the alleged relationship using cross-section data at the SMSA level. She found her anticipated connection to be verified statistically, but the relevance of her work for black families is questionable. In particular, black males do not typically earn the mean wage received by all male workers in the manufacturing sector. Moreover, throughout the 1970's, AFDC payments did not continue to rise relative to

the manufacturing wage, but female headship grew explosively among blacks. In addition, there were other problems with her study that carry over to subsequent work. Among these is the fact that she restricted her attention to female-headed families formed by divorce or separation.

Ross and Sawhill also tested the same model using cross-section data from forty-one U.S. cities. They, like Honig, focus solely on the growth in female headship associated with marital dissolution. This is especially important because it omits a major source of the increase in female-headed families among blacks. By 1979, the largest share of black families headed by women consisted of single, never-married women. This means, in turn, that a growing share of black female-headed families are formed either when mothers have and keep children newly born out of wedlock and set up separate households, or set up separate households after having lived with relatives for some period of time.

A similar difficulty plagues a study by Kathy Bradbury et al. where the findings on black female headship were mitigated by the complete inattention devoted to families of never-married black women. They confined their research to women heads in the 25-34 year age range, thus lopping off younger women who constitute a significant part of the growth in female-headed families.

To the extent that all these studies opt for a model that treats female headship as the outcome of an optimum choice subject to constraints by mothers, they all share an odd common characteristic. They ignore the potential constraint set by the sheer availability of men as potential spouses. This is especially evident in the paper by Danziger et al. which assumes an unlimited supply of men. Where black women are concerned, this is an odd assumption given fairly forceful expressions from black women themselves that there is a shortage of black males. Furthermore, forty years ago, sociologist Oliver Cox called explicit attention to the importance of the sex ratio and the incidence of female headship among blacks. Cox contended that black women were confronted with a situation where the black male population is *both* less "economically able to

marry" and is less available numerically than white males for white females.

The reliance of the contemporary studies on cross-section data without any attempt to account for "rational" migration prevents examination of a further implication of the new economics of the family. These Beckerian models imply a rapid reversal would occur in family structures if economic conditions changed in favor of male earnings relative to female earning. But whether or not such reversibility is occurring is more readily examined with time-series data.

Also we note that our own efforts to perform a preliminary Granger-Sims statistical "causality" test, reported in detail in our 1982 paper, suggests that the feedback loop runs from female-headed households onto AFDC payments rather than vice versa. We are aware of criticisms of the Granger-Sims concept of causality. Nevertheless, we mention these results as a further indicator of the inadequacy of existing empirical research based upon the economic model.

A final clue that something might be amiss with the economic model of family structure determination is well known ever since the results of the income maintenance experiments have been made available. Its full implications simply have not been pursued. Public assistance provided under a negative income tax scheme to two-parent families without the AFDC absent-father means test did not reduce the incidence of marital splits. John Bishop has reported that in "three of the four [sites where the experiments were performed], the measured rates of marital dissolution were larger in the experimental [NIT] group than in the control [AFDC] group" (p. 312). Also Bishop (pp. 310-11) has noted that several states have a program that provides cash assistance to two-parent families when the head is unemployed (AFDC-UP); limited experience with those programs indicated that marital splits rose rather than fell after their inauguration.

### III. A New Test of the Economic Model

Due to the problems described in the previous section, we reconstruct the family structure model rooted in the Becker ap-

proach and reestimate an equation for the determinants of black family female headship with annual time-series data for the period 1955-80. We propose the following simultaneous equation model:

- (1)  $BFFH = f_a(AFDC, BAFDC, ILLEG, MORT, RATE\ 1, AGE5, AGE6, AGE7, MF),$
- (2)  $BAFDC = f_b(MF, BFFH, AFDC, ILLEG, MORT, RATE\ 3, AGE5, AGE6, AGE7),$
- (3)  $AFDC = f_c(BFFH, BAFDC, ILLEG, MORT, RATE\ 3, AGE5, AGE6, AGE7).$

The term *BFFH* is the odds in favor of a black family being female headed, or the ratio of black female family householders to all other black families in a given year. The term *AFDC* is defined as  $A/[(1-U)Y/12]$  where *A* is the mean monthly AFDC payments per AFDC family. *U* is the mean nonwhite male unemployment rate, and *Y* is the mean earning of full-time employed nonwhite males in a given year. The term *BAFDC* is the percentage of nonwhite families on welfare; *ILLEG* is the ratio of illegitimate nonwhite births to nonwhite females; *MORT* is the nonwhite male death rate; *MF* is the ratio of mean earnings for nonwhite males to nonwhite females for year-round full-time workers; *RATE 1* is the ratio of black females 18 years and older to black males above 18; *RATE 3* is the ratio of black females to males from 20 years up to 39 years of age; *AGE6* is the percentage of black females 20 to 29 years of age; *AGE7* is the percentage of black females 30 to 39 years of age; and *AGE5* is the percentage of black females older than 65.

We explain in detail our data sources and our procedures for constructing variables in our earlier paper. Wherever possible we attempted to refine the data to enable us to use

statistics on blacks rather than all nonwhites. This proved to be especially important in the construction of the time-series for black female to male ratios.

From the simultaneous equation system (1) to (3), we report here the results for estimation of equation (1), estimated in *log*-linear form with a constant term. Full results for estimation of the complete system also are available in our earlier paper. We estimated equation (1) using an instrumental variable technique, using all exogenous variables plus a linear time trend, the nonwhite male unemployment rate and the ratio of black females to males in the age group 18–65 years as instruments, to correct for simultaneous equation bias. The following results emerge:

$$\begin{aligned} \ln(BFFH) = & 2.77 - .34 \ln(AFDC) \\ & (-.84) \\ & - .07 \ln(BAFDC) + 6.53 \ln(RATE\ 1) \\ & (-.39) \quad (1.88) \\ & + 2.34 \ln(AGE6) + 1.00 \ln(AGE7) \\ & (3.05) \quad (1.94) \\ & + 1.64 \ln(AGE5) - .17 \ln(ILLEG) \\ & (1.06) \quad (-.29) \\ & + 1.02 \ln(MORT) + .29 \ln(MF), \\ & (1.15) \quad (.37) \end{aligned}$$

$$SEE = .0519; \quad D.W. = 2.20; \quad N = 25.$$

Our results indicate, in stark contrast with the impression given by the studies reviewed above, that the economic variable *AFDC* does not play an important role in *statistically* explaining the growth in female headship. The extent of welfare dependency measured by *BAFDC* similarly fails to play a more important role. Further, relative nonwhite male-female earnings show no effect on female headship. In our complete model estimates, we find that *MF* does seem to have an impact on *BAFDC*. Dropping *MF* from the regression leaves all coefficients amazingly stable (and significantly lowers

the standard error on the important variable *MORT*).

What matters the most here are the demographic variables. They alone have *t*-statistics providing statistical significance at the 95 percent confidence level. Variations in the black female to male ratio, *RATE* 1, displays the largest absolute effect on the incidence of female headship. The variable with the estimated coefficient next in magnitude of importance is the proportion of black females between 20 and 29 years of age. The coefficient *AGE*7, for the cohort 30 to 39, also is statistically significant. The U.S. Census (1980) tabulations demonstrate both that there are more black women in the population "at risk" of being single parent family heads and the within-cohort probability of a black woman being a female family householder has increased perceptibly over the decade.

The sign of the coefficient of the mortality variable reinforces the conclusion that the relative lack of black men contributes to the growth in black female headship. Deaths among black males, especially young men in their late teens and early twenties, depletes the pool of potential husbands. The growing proportion of black women in the marriageable years 20 to 29, could contribute to growth in female headship because of the social processes working to remove black males as potential marriage partners. They also could contribute to the growth in female headship because of the *cumulative consequences* of an intergenerational transmission mechanism as daughters from female-headed homes now become adults and set up their own female-headed families. Relevant here also might be the possible unspoken issue of social class. Black professional women, a distinct minority of all black women, appear to be living as singles without children in far greater proportions than low-income black women, who thus tend to carry on the bulk of black family life.

Regardless, the image that emerges here is far different from the one that envisions black women responding to perceived marginal benefits in income and "leisure" causing them to select family arrangements without husbands. Furthermore, these results

believe the conclusion that alterations in the welfare system might reverse the momentum now generating such a high ratio of female headed to male headed families among blacks.

Black female headship and black family poverty are better understood as two symptoms of deeper historical processes that relate to the ongoing dependent status of blacks in America. Disproportionately located at the bottom of American society, black families long have been among the most vulnerable and hence most adversely affected by major social upheavals. Without institutional protections from within the black community, blacks have had to rely on external protections from sources whose own interest lay elsewhere. Thus welfare dependency is only the most recent signal of the continued lack of self-determination for black families in America.

#### REFERENCES

- Becker, Gary S., *A Treatise on the Family*, New York, 1981.
- Bishop, John H., "Jobs, Cash Transfers, and Marital Instability: A Review and Synthesis of the Evidence," *Journal of Human Resources*, Summer 1980, 15, 312-21.
- Bradbury, Kathy et al., "Public Assistance, Female-Headship and Economic Well-Being," *Journal of Marriage and the Family*, 1979, 41, 519-35.
- Coe, Richard, "Welfare Dependency: Fact or Myth?," *Challenge*, September/October 1982, 25, 43-49.
- Cox, Oliver, "Sex Ratios and Marital Status Among Negroes," *American Sociological Review*, 1940, 5, 937-47.
- Danziger, Sheldon et al., "Work and Welfare as Determinants of Female Poverty and Household Headship," *Quarterly Journal of Economics*, August 1982, 97, 519-34.
- Darity, William, Jr., "The Class Character of the Black Community: Polarization Between the Black Managerial Elite and the Black Underclass," *Black Law Journal*, Fall 1981, 7, 21-31.
- \_\_\_\_\_ and Myers, Samuel L., Jr., "Black Family Structure and Welfare Dependency: Details of the Statistical Investigation," unpublished manuscript, University of Pittsburgh, December 1982.
- Frazier, E. Franklin, *The Negro Family in the United States*, Chicago: University of Chicago Press, 1939.
- Honig, Marjorie, "AFDC Income, Recipient Rates, and Family Dissolution," *Journal of Human Resources*, Summer 1974, 9, 303-22.
- Miller, Loren, *The Petitioners: The Story of the Supreme Court of the United States and the Negro*, New York, 1966.
- Moynihan, Daniel, "The Negro Family: The Case for National Action," in Lee Rainwater and William Yancey, eds., *The Moynihan Report and the Politics of Controversy*, Cambridge, 1967.
- Rein, Martin and Rainwater, Lee, "How Large is the Welfare Class?" *Challenge*, September/October 1977, 20, 20-23.
- Ross, Heather and Sawhill, Isabel, *Time of Transition: The Growth of Families Headed by Women*, Washington, 1975.
- U.S. Department of Commerce, Bureau of the Census, "Families Maintained by Female Householders 1970-1979," *Current Population Reports*, Special Studies Series P-23, October 1980.

# Budget Cuts as Welfare Reform

By SHELDON DANZIGER\*

President Nixon's 1969 proposal for a Family Assistance Plan (FAP) established welfare reform as a major social policy goal. In his first year in office, President Carter also placed welfare reform—the Program for Better Jobs and Income (PBJI)—high on his legislative agenda. The FAP and PBJI were both variants of the negative income tax. They shared several common elements, including a national minimum income guarantee, an extension of benefits to persons who were categorically ineligible under existing programs, a concern with maintaining work incentives by keeping marginal benefit reduction rates on earnings well below 100 percent, and a belief that a reformed welfare system could control the growth in costs and case loads. Both also generated fatal political opposition and harsh criticism from welfare analysts who pointed out that the “iron triangle”—the tradeoffs among income guarantees, work incentives, and total costs—made their goals mutually inconsistent.

President Reagan also placed welfare reform at the center of his social policy agenda in his first year in office. Unlike his predecessors' plans, his reform was successful. By October 1981, a “drastic fiscal retrenchment” had been proposed, legislated, and implemented in the largest cash welfare program, Aid to Families with Dependent Children (AFDC). The Reagan reform, incorporated in the Omnibus Budget Reconciliation Act of 1981 (OBRA), does not confront the iron triangle of negative income tax plans like FAP or PBJI. It does not attempt to reduce poverty by altering income guarantees or

extending eligibility. It does not attempt to encourage work effort by lowering marginal tax rates on recipients. Supply-side logic notwithstanding, it reduces costs and case loads by raising the tax rate on welfare recipients' earnings to 100 percent and by establishing more restrictive gross income limits.

President Reagan has reformed welfare by cutting the budget. He has clearly reduced welfare dependency in the short run, as the number of AFDC recipients has declined in most states by between 10 and 15 percent in the last year. A complete evaluation of the long-run effects of the OBRA reforms on the economic well-being and work effort of welfare recipients must await data on behavioral responses that have only recently been induced. Nonetheless, an analysis of the redistributive effects of welfare in recent years can provide a basis for estimating how reduced welfare dependency will affect economic well-being in the short run.

## I. Poverty and Dependency on Income Transfers

That the Reagan tax and budget program will increase poverty and inequality in the near term is generally accepted. In addition to the analyses of the Congressional Budget Office and John Palmer and Isabel Sawhill, Michael Boskin, and other supporters of the budget cuts acknowledge these effects. This paper focuses on the welfare cuts and suggests how their effects will differ by race.

Table 1 presents 1980 data on the incidences of pretransfer and official (after cash transfers) poverty of white and nonwhite persons; on the dependence of the pretransfer poor on cash welfare and nonwelfare transfers, measured by the percentage of persons living in households that receive these transfers; and on the antipoverty effectiveness of the two types of transfers, measured by the percentage of pretransfer poor persons taken out of poverty by transfers. Persons are further classified by the age and sex

\*Institute for Research on Poverty, University of Wisconsin-Madison. This research was supported by grants from the Graduate School Research Committee of the University of Wisconsin-Madison and the Alfred P. Sloan Foundation. Sally Davies provided valuable assistance. S. Cole, R. Haveman, G. Jakubson, R. Plotnick, and E. Smolensky offered helpful comments on an earlier version.

TABLE 1—POVERTY AND DEPENDENCY ON CASH TRANSFERS, 1980

Head of Household <sup>a</sup>	Incidence of Pretransfer Poverty (1)	Percentage of Pretransfer Poor Persons:				
		Receiving Cash Welfare <sup>b</sup> (2)	Taken Out of Poverty by Cash Welfare (3)	Receiving Nonwelfare Cash Transfers <sup>c</sup> (4)	Taken Out of Poverty by Cash Non- welfare Transfers (5)	Official Incidence of Poverty (6)
White Nonaged						
Male	9.8%	20.4%	4.9%	47.1%	26.4%	6.8%
Female	35.9	46.0	6.5	33.4	15.8	27.9
White Aged						
Male	49.3	9.1	1.8	97.7	81.6	8.2
Female	67.9	16.5	4.1	95.9	60.1	24.3
Nonwhite Nonaged						
Male	21.0	38.2	7.2	42.1	14.3	16.5
Female	59.3	68.0	6.5	28.8	7.2	51.2
Nonwhite Aged						
Male	67.2	31.7	8.3	93.3	50.6	27.6
Female	83.2	54.9	12.3	85.9	27.8	49.9
All Persons	21.9	30.9	5.1	58.2	35.3	13.0

Source: Computations by author from March 1981 *Current Population Survey*.

<sup>a</sup>Heads of households 64 years of age or younger are nonaged; those 65 or older are aged.

<sup>b</sup>Cash welfare transfers include AFDC, Supplemental Security Income, and General Assistance.

<sup>c</sup>Nonwelfare cash transfers include Social Security, Railroad Retirement, Unemployment Compensation, Worker's Compensation, Government Employee Pensions, and Veterans' Pensions and Compensation.

of the head of their household. The data reflect the well-known large differences in poverty between majority and minority, between male-headed and female-headed, and between nonaged and aged households. The incidence of pretransfer poverty was 21.9 percent for all persons, ranging from 9.8 percent for those headed by nonaged white males to 83.2 percent for those headed by aged nonwhite females.<sup>1</sup>

The pretransfer poor are highly dependent on both welfare and nonwelfare transfers. However, only 30 percent of the pretransfer poor received welfare. Nonwhites are more likely to receive welfare and less likely to receive nonwelfare transfers than are whites. Although the large and increasing expenditures on income maintenance programs have

been a topic of great concern, less attention has been focused on the gaps in coverage in the present system—the holes in the safety net. Almost 40 percent of nonaged, poor households receive no income transfers, and many of those who do receive transfers do not receive enough to lift their households above the poverty line. Much of the variation in coverage among the poor is due to the different eligibility requirements and benefit levels in programs administered by the states.

In this respect, the Reagan reforms differ sharply from FAP or PBJI. While those plans would have established a national minimum benefit, the proposed "New Federalism" would provide the states with more discretion. And Howard Chernick's study of the incentive effects of block grants suggests that this proposal would reduce incentives for states to maintain existing benefits.

Cash transfers reduce poverty by 40 percent, from 21.9 percent to the officially reported 13.0 percent. Most of the transfer reciprocity and poverty reduction is accounted for by nonwelfare transfers. Holding age and sex of head constant, poor nonwhites are

<sup>1</sup>Pretransfer income is calculated by subtracting government transfers from posttransfer income. While this definition assumes that transfers elicit no behavioral responses, transfers do induce labor supply reductions. As a result, recipients' net incomes are not increased by the full amount of the transfer and the pre/post comparisons made here will provide upper-bound estimates of the antipoverty effects of transfers.

less likely to be removed from poverty by all cash transfers for each of the groups shown in Table 1. Cash welfare benefits, however, have a bigger impact for nonwhites than for whites.

The Reagan budget cuts have left Supplemental Security Income and Social Security largely untouched (as of this date). As a result, poverty among the aged has not changed significantly. Among the nonaged, the income guarantees of nonworking welfare recipients have not been reduced. In 1980, roughly 90 million persons, over 40 percent of all persons, lived in households receiving some type of cash transfer. My estimate is that about 5 million persons—a relatively small percentage of all transfer recipients—will experience substantial income reductions because of the welfare cuts. This group, however, will be disproportionately nonwhite and female. For example, persons living in households headed by nonaged, nonwhite women are only 5.4 percent of all persons living in households headed by the nonaged, but they compose 19.1 percent of the pretransfer poor and 34.0 percent of the pretransfer poor who receive cash welfare. Even before the Reagan cuts, they had a higher official incidence of poverty than any of the other groups in Table 1.

Some analysts (see, for example, George Gilder) have claimed that those whose welfare dependency has been terminated will, in the long run, have more motivation, self-esteem, and success in the labor market. Nonetheless, in the short run they will be more likely to remain poor. In the next section, I estimate this short-run effect.

## II. The Reagan Reform and the Economic Well-Being of AFDC Recipients

Two of the many AFDC changes typify the Reagan Administration's rejection of the negative income tax. The first is the introduction of an income "notch"—a recipient is no longer eligible for benefits if gross income exceeds 150 percent of the state's need standard. The second is that after four months of earnings, the marginal benefit reduction rate increases to 100 percent. Under prior law, the first \$30 of earnings were not

taxed, and the remainder were taxed at a nominal rate of 67 percent.

The magnitude of the effects of these changes on work effort cannot be estimated at the present time. However, the direction of the effects can be inferred. About half of the AFDC recipients in the March 1981 *Current Population Survey (CPS)* reported that they had not worked at all during 1980. Because welfare guarantees are not changed and the tax rate on earnings is increased, they will have a reduced incentive to begin work. Another quarter of the recipients have earnings that exceed their welfare guarantees. When they are removed from the welfare rolls, their tax rate will also fall, so both the income and substitution effects will lead to increased work incentives. However, if additional work is not available, depending on the value placed on leisure, they may have an incentive to reduce work effort in order to return to welfare. The remaining quarter have yearly earnings (usually below \$3,000) that are lower than their welfare guarantees. In the short run, they will be no worse off if they quit working. While the net effects of the cuts on work effort are ambiguous, the disincentives are likely to predominate at current unemployment rates.

The short-run effects on family income can be estimated from data on earnings, welfare benefits, other incomes, work and child care expenses and family size. Because the *CPS* does not report expenses, I use a July 1981 sample of about 4,500 AFDC cases drawn at random from the State of Wisconsin's computerized administrative records (see Sally Davies).

Table 2 shows the actual economic status of these cases before and the simulated status after the OBRA reforms, on the assumption that working recipients had been on the rolls for four months, and that labor supply and work expenses remained constant.<sup>2</sup> The case load is divided into four groups as shown in columns 1–4. About 9 percent of the case load is estimated to have been terminated

<sup>2</sup>The reduced AFDC benefits imply reduced Medicaid eligibility, and increased Food Stamp benefits. Medicaid benefits were not available, and the change in Food Stamps was not simulated.

TABLE 2—ESTIMATED EFFECTS OF OBRA REFORMS ON ECONOMIC STATUS

Case Characteristics	Terminated Because of Gross Income Levels (1)	Terminated Because of Increased Tax Rate (2)	Benefits Reduced Because of Increased Tax Rate (3)	No Change (4)	All Cases (5)
Distribution of Case Load	4.5%	4.7%	11.4%	79.4%	100.0%
Gross Monthly Earnings	\$915	\$674	\$362	\$0	\$114
Monthly Earnings Less Work and Child Care Expenses	\$519	\$450	\$216	\$0	\$68
Pre-OBRA AFDC Benefit	\$226	\$233	\$383	\$404	\$386
Post-OBRA AFDC Benefit	\$0	\$0	\$220	\$404	\$346
Percentage Reduction in Disposable Income <sup>a</sup>	-30.4%	-34.1%	-27.2%	0.0%	-9.3%
Poverty Incidence Pre-OBRA <sup>b</sup>	0.0%	0.0%	28.1%	100.0%	82.5%
Poverty Incidence Post-OBRA <sup>b</sup>	0.0%	13.2%	63.3%	100.0%	87.1%
Mean Persons per Case	2.4	2.6	2.9	2.7	2.7

Source: Computations by author from July 1981 sample of Wisconsin AFDC cases.

<sup>a</sup>Defined as reduction in AFDC benefit plus earnings less work expenses less child care.

<sup>b</sup>As officially measured, these computations do not include the values of Food Stamps, Medicaid or other in-kind assistance.

because of the 150 percent of needs standard gross income limit and the increased tax rate; 11 percent to have reduced benefits because of the increased tax rate. Four-fifths of the case load was not working in July 1981 and was thus unaffected.<sup>3</sup>

The OBRA changes reduce the disposable incomes of the average AFDC recipient by 9 percent and increase poverty as officially measured from 82.5 to 87.1 percent. What is especially striking is that poverty among recipients in a high-benefit state like Wisconsin was so widespread before OBRA. Because the distribution of work effort is so skewed, the averages for all cases obscure very different patterns. All of the cases where the head did not work were poor before the reforms. The reforms reduce disposable income by about 30 percent for cases with earnings, and significantly increase poverty for the two groups affected by the increased tax rate.

Particularly hard hit are the 11.4 percent of the cases where earnings are low enough

so that eligibility for a reduced benefit is maintained. Poverty for this group doubles, from less than 30 to more than 60 percent. These recipients face a strong work disincentive, since their average disposable income after an average of 24 hours of work per week (\$436) is only slightly higher than that of nonworking recipients (\$404). Poverty increases from zero to about 13 percent for cases terminated because of the higher tax rate (col. 2). Reinstatement of the \$30 and one-third income disregard would offset many of the poverty-increasing effects of the OBRA reform, but would cut budgetary savings by over one-half.

The 4.5 percent of the case load terminated because of the gross income limit (col. 1) provides a contrast. By federal poverty standards they are not truly needy, as none are poor either before or after OBRA. They were eligible under prior rules because they reported work and child care expenses that averaged over 40 percent of their earnings. In Wisconsin, where 150 percent of the needs standard is well above the poverty line, this change reduces costs and case loads without increasing poverty. In many states, however,

<sup>3</sup>While only 20 percent of the case load was working in a given month, a much larger percentage works at some point during the year.

150 percent of the needs standard is well below the poverty line, and the gross income limit increases poverty. If this income notch were set instead at the federal poverty line, the Wisconsin results would generally hold across the nation.

The Wisconsin data show that the Reagan welfare cuts have reduced the number of welfare recipients removed from poverty by cash transfers. Their direct effect on poverty among all persons has been small because welfare recipients are a minority of all transfer recipients and because only a minority of welfare recipients—those with earnings—have been significantly affected by the reforms implemented thus far. But for a majority of recipients with earnings, the effects have been substantial.

The data in Table 1 showed that nonwhites were not more likely to be dependent on cash transfers than whites, but that they were more likely to be dependent on welfare. For example, in 1981 about 8 percent of all nonwhite children and 1.5 percent of all white children lived with mothers who both received welfare and worked. If the increased poverty in Wisconsin is representative of the national effect, then nonwhites will be disproportionately affected by the AFDC cuts.

### III. Policy Implications

I have shown that current transfer programs significantly reduce poverty, but that welfare accounts for only a small proportion of this reduction. Increased welfare, however, cannot reduce poverty, increase work effort, and contain the welfare rolls. One promising alternative would be a targeted employment program that allows recipients to mix work and welfare. From this perspective, the major flaw of OBRA is that it requires recipients to choose either work or welfare. As a result, those not currently working, who would have gradually increased their earnings and reduced their welfare benefits under prior law, will be less likely to begin to work.

The Supported Work Demonstration project (Manpower Demonstration Research Corporation) provides an example of both the difficulty of reducing case loads and the antipoverty possibilities of a targeted em-

ployment program. George Jakubson and I used Supported Work data to simulate the national effects of implementing such a program. We found that over 80 percent of the AFDC participants would have been poor if they merely had access to current transfer programs (quite similar to the Wisconsin poverty incidence in Table 2), whereas only 35 percent would have been poor if they also had access to the jobs program. Unlike a negative income tax, such a program increases work effort and reduces poverty. But, total program costs would have increased significantly, and welfare would still have accounted for over a quarter of recipient incomes.

President Reagan's welfare reform has reduced AFDC case loads and increased poverty for many welfare recipients who were mixing work and welfare. The Reagan program assumes that those who remain poor will be better off waiting for economic growth to trickle down from those above them rather than relying on welfare and public jobs programs. However, given the recent projections of high unemployment and slow economic growth through the mid-1980's, it is likely that the wait confronting these families will prove to be longer than they or the Reagan Administration anticipate.

### REFERENCES

- Boskin, Michael, "Reaganomics and Income Distribution: A Longer-Term Perspective," *Journal of Contemporary Studies*, Summer 1982, 5, 31-44.
- Chernick, Howard, "Block Grants for the Needy," *Journal of Policy Analysis and Management*, Winter 1982, 1, 209-22.
- Danziger, Sheldon and Jakubson, George, "The Distributional Impact of Targeted Public Employment Programs," in R. Haveman, ed., *Public Finance and Public Employment*, Detroit: Wayne State University Press, 1982.
- Davies, Sally, "The Effects of the Omnibus Budget Reconciliation Act on the Well-Being of AFDC Recipients in Wisconsin," discussion paper, Institute for Research on Poverty, University of Wisconsin, Decem-

ber 1982.

Gilder, George, *Wealth and Poverty*, New York: Basic Books, 1981.

Palmer, John, and Sawhill, Isabel, *The Reagan Experiment*, Washington: Urban Institute, 1982.

Manpower Demonstration Research Corporation,

*Summary and Findings of the Supported Work Demonstration*, Cambridge: Ballinger, 1980.

U.S. Congressional Budget Office, "Effects of Tax and Benefit Reductions Enacted in 1981 for Households in Different Income Categories," Special Study, February 1982.

## INVESTMENT, SAVINGS, AND INCENTIVES

### The Determinants of Investment: Another Look

By BEN S. BERNANKE\*

Studies of the determinants of business fixed investment have typically (although not universally) emphasized output, sales, or profits variables over measures of capital costs. For example, in a study that compared leading econometric models of investment, Peter Clark concluded that "output is clearly the primary determinant of nonresidential fixed investment" while, at least in the short run, "the effect of moderate variations in taxes and interest rates is likely to be negligible..." (1979, pp. 103-04).

However, the idea that capital costs are not important in the short run has recently been battered by events. Since the 1979 regime change at the Federal Reserve, real interest rates have been unusually high. These high rates have been widely blamed for the fact that, despite the political commitment to support capital formation, investment spending has been weak. In this view, the two recent recessions are a result, rather than a cause, of the low demand for capital goods (and other durables).

With an eye toward measuring the effect of interest rates, this note reexamines the determinants of nonresidential investment. The model underlying my results is of the linear-quadratic variety and is closely related to the labor demand model of Thomas Sargent (1978). This model is fairly restrictive, which may prevent it from being useful in some contexts. Its virtue is that it permits estimation to be based on the closed-form solution to a dynamic stochastic optimization problem, which leads to maximum efficiency in the use of the data. The estimation procedure employed here is not vulnerable, as those in some earlier studies are, to the criticisms made by Robert Lucas (1976).

Using data up to 1979, I find the model to be largely, although not entirely, successful as a description of U.S. aggregate investment. However, the importance of capital costs for investment comes through strongly. It is plausible to conclude that high real interest rates are a major source of the recent sluggishness in capital expenditure.

#### I. Model Setup

In the model the representative investing firm is assumed to maximize an objective function  $V_t$ , given by

$$(1) \quad V_t = C_t + E_t \left\{ \sum_{i=1}^{\infty} \left( \prod_{j=0}^{i-1} (1 + r_{t+j})^{-1} \right) C_{t+i} \right\}.$$

The cash flow in period  $t$ ,  $C_t$ , is defined by

$$(2) \quad C_t = a_t K_t - J_t (K_{t+1} - (1 - \delta_t) K_t) - (d/2)(K_{t+1} - K_t)^2$$

where  $a_t$  = returns to fixed capital stock (in period  $t$ );  $K_t$  = capital stock (at the beginning of  $t$ );  $J_t$  = real after-tax price of new capital goods;  $\delta_t$  = physical depreciation rate of capital;  $r_t$  = prospective real interest rate; and the  $C_{t+i}$  are similarly defined.  $a_t$ ,  $J_t$ , and  $\delta_t$  are realizations of random variables that are known at the beginning of  $t$ . The real interest rate in  $t$  is also random but its realization is not known until  $t + 1$ . The random variables may have any of a broad class of distributions, as long as a particular transversality condition on the valuation of capital is satisfied. The parameter  $d$  reflects internal costs of adjustment.

The multiplicative "production function" used here is simple, but it is not as restrictive

\*Stanford University

as it may appear. Sufficient conditions for the total return to capital to be of the form  $a_t K_t$  are 1) there are constant returns to scale and 2) capital is the only quasi-fixed factor of production. With these assumptions the random variable  $a_t$  can be thought of as embodying optimal short-run utilization and variable input decisions. The constant returns to scale assumption is consistent with the data; my measure of average returns to capital shows no trend or sensitivity to the absolute level of capital stock.

The objective function specified above departs from the usual linear-quadratic formulation in permitting a stochastic and variable real interest rate. The usual  $LQ$  model assumes a constant interest rate because only in that case are exact closed-form decision rules obtainable. The obvious importance of permitting a variable interest rate in this study leads us to take a different approach, based on an approximation to the exact solution. Employing a method used by Andrew Abel and Olivier Blanchard (1982) in a somewhat different context, I begin by taking a second-order Taylor expansion of the realized value of  $V_t$  around a "long-run" discount factor  $(1+r^*)^{-1}$  and the current capital stock  $K_t$ . I then take expectations and differentiate with respect to the control variables (the future capital stocks). The stochastic difference equation that satisfies the resulting first-order conditions is easily found and may be thought of as representing an approximate solution rule for the original problem. For period  $t$  that solution is

$$(2) \quad K_{t+1} - K_t = (1/d) E_t \sum_{i=0}^{\infty} (1+r^*)^{-i} \\ \times \{ (1+r_{t+i})^{-1} a_{t+i+1} \\ - (J_{t+i} - (1+r_{t+i})^{-1} J_{t+i+1} (1-\delta_{t+i+1})) \}$$

Setting  $r_{t+i} = r^*$  in (2) yields the exact solution for the constant interest rate case. Equation (2) can be rewritten as

$$(3) \quad K_{t+1} - K_t = (1/d) E_t \sum_{i=0}^{\infty} (1+r^*)^{-i} \\ \times \{ (1+r_{t+i})^{-1} RET_{t+i+1} - COST_{t+i} \}$$

where  $RET$  (equal to the random variable  $a$ ) is a measure of the gross return to capital and  $COST$  is the conventionally measured one-period holding cost of capital. Equation (3) says that net investment is proportional (approximately) to the present value of expected net returns to capital, with the adjustment cost parameter  $d$  determining the factor of proportionality.

## II. Data

The above model was applied to annual U.S. aggregate data, 1947–79. Separate equations were estimated for equipment and non-residential structures, which required the additional assumptions that equipment and structures are separable in production and have proportional returns. These are extremely strong assumptions; however, it seemed better to make them than to ignore the potentially large difference between equipment and structures in the size of adjustment costs. (Separate equipment and structures equations are in fact quite common in the investment literature.)

The constant-dollar net capital stock series were taken from the study by John Musgrave (1981). These series end in 1979, which explains my choice of sample endpoint. Implicit depreciation rates  $\delta_t$  were calculated using these series and gross capital expenditure data from the *Survey of Current Business*. I divided the net investment series by population as a means of removing the trend. Net real investment per capita in 1979 was about \$200 (1972 dollars); about 65 percent of this was equipment.

Gross returns to capital were found as follows: Total gross capital income ( $CAPINC$ ) was defined to be the sum of real profits, depreciation, and interest. With the assumption that the gross return to structure in each period is  $x$  times the return to equipment, we can write

$$(4) \quad RETEQ = CAPINC / (KEQ + xKSTR)$$

$$(5) \quad RETSTR = xRETEQ$$

where  $RETEQ$  and  $RETSTR$  are the returns to equipment and structures, respectively, in a given period, and  $KEQ$  and  $KSTR$  are

equipment and structures stocks. Imposing the requirement that long-run average net returns to equipment and structures be equal, I found  $x = .785$  in my data. Values of  $x$  in this range were also found by alternative methods.

The returns variables are procyclical but essentially trendless. It is interesting that, by this measure, gross returns to capital were at record highs in 1976–80. These remarks continue to apply (and the results of the estimation are unchanged) when capital income is defined to exclude a generous estimate of real interest paid to bondholders. (However, total exclusion of interest from capital income significantly worsened the performance of the returns variables in the estimation.)

To construct series for holding costs, data were required on the real prices of capital goods, taxes, and interest rates. For capital goods prices, I used Musgrave's deflators, divided by the GNP deflator. The federal government subsidizes investment via tax credits and depreciation deductions; the effective subsidies to investment in equipment and in structures in each year were found by extending the calculations of Robert Hall and Dale Jorgenson (1967). Future depreciation deductions were assumed to be discounted by the Baa corporate bond rate. The weighted average of my subsidy series for equipment and for structures corresponded fairly closely to the series for all investment computed by Lawrence Summers (1981).

I tried a number of interest rates in the calculation of capital holding costs, without finding qualitatively significant differences. Since the theory actually calls for a short-term rate, reported estimates used the six-months commercial paper rate. The long-run average real rate,  $r^*$ , was assumed to be .02; the results were not very sensitive to this assumption.

I wanted an *ex ante*, rather than an *ex post* measure of capital holding costs, since it is the former that affects firm decisions. This required information on expected as well as realized values of the key variables. A difficult problem was modelling investor anticipation of changes in the tax law. In an illustrative example, Lucas treated the investment tax credit as an exogenously determined Markov process. This will not work

in applications, since there is a large endogenous component to investment tax laws: Investment subsidies are regularly increased at troughs and decreased at peaks of the investment cycle. My approach was to take the next-period expectation of the total tax break for investment to be the fitted value of a regression of the subsidy on lagged subsidies and lagged investment. (Lagged investment turns out to be an important and statistically significant determinant of current investment tax laws.) Other expectations—of capital goods prices and the general price level for the next period—were similarly modelled by prediction equations.

The holding costs series so constructed took on their highest values in 1958–63 and 1970–72, their lowest values in 1964–69 and 1973–74. Since 1975 costs have been at intermediate levels for equipment, a bit higher for structures.

### III. Estimates

The estimation method followed Sargent. This approach requires that the time-series properties of the explanatory variables be modelled. Experimentation showed that the returns variable and two costs variables could be represented reasonably well as *AR1* processes plus constants; although more general *AR* processes could have been handled, this made estimation particularly simple. The three autoregressions were estimated simultaneously with equipment and structures equations based on equation (3). The results were

Parameter	Estimate	( <i>t</i> -statistic)
<i>R1</i>	0.83	(23.90)
<i>R2</i>	0.76	(20.05)
<i>R3</i>	0.65	(10.51)
<i>DE</i>	.00209	(4.70)
<i>DS</i>	.00651	(4.02)

where *R1*, *R2*, and *R3* are autoregressive coefficients for returns to capital, equipment costs, and structures costs, respectively; and *DE* and *DS* are the adjustment cost parameters for equipment and for structures. Estimates of constants are not reported. *AR1* corrections were made in the equipment and structures equations using autoregressive parameters found in a first-stage regression;

the *AR* parameter values were 0.7513 and 0.8526, respectively. Separate estimation of the equations yielded similar results.

As will be reported, these estimates imply significant effects of capital costs on the demand for investment goods. It might properly be objected, however, that the restrictive form of the estimated equations imposes this answer in advance. To examine this possibility, I reestimated the model without restrictions on the coefficients of the returns and costs variables. For equipment, it was found that the unrestricted estimates were very close to those implied by the restricted model. For structures, the costs coefficient was large and sharply estimated, but the estimated sensitivity to the returns variable was small and insignificantly different from zero. Although the structures part of the model could not be statistically rejected with 90 percent confidence, it is clear that this component of investment is not described as well by the adjustment-costs model as is expenditure on equipment. (This is not surprising, given that delivery lags for structures are typically longer than the interval of observation in the sample.)

The unrestricted equations were also estimated by an instrumental variables technique (Fair's method) to reduce simultaneity bias. The costs variables again came through strongly as explanators of investment spending. It does not appear, then, that the estimated importance of costs is an artifact of the restricted model.

Using the estimates from the complete model, I performed a variety of *ceteris paribus* simulation experiments (using 1979 as a base year.) It was assumed in each case that investors expected innovations in returns or costs to display the same persistence as observed in the sample. The flavor of the results may be suggested by the following:

A one percentage point innovation in the gross returns to capital (a sustained economic recovery usually adds three to four points to gross returns) increases net investment in equipment by 16.2 percent and in structures by 7.7 percent during the first year.

A one percentage point innovation in the investment tax credit raises net equipment

investment 1.9 percent and net structures investment 0.3 percent in the first year.

A one percentage point innovation in the real interest rate (nominal rate held constant), reduces net equipment investment by 12.1 percent and net structures investment by 6.3 percent in the first year.

These experiments compare effects of certain variables on the (partial equilibrium) investment demand schedule. In this sense, the marginal revenue product of capital, as in earlier studies, is seen to be important (at least for equipment). However, real interest rates, acting both through the cost of capital and the discount factor for future returns, are also a powerful influence. Given these results, it is unsurprising that recent tax legislation favorable to investment has been unable to offset the depressing effect of high interest rates.

I found experimentally that investment prediction equations which used both the variables from this study and also financial market variables (like Summers' "tax-adjusted *q*") significantly outperformed equations based only on one set of variables or the other. In ongoing research using panel data, I am attempting to augment the "fundamentalist" approach of this paper with information drawn from market valuations of capital.

## REFERENCES

- Abel, Andrew B. and Blanchard, Olivier, J., "The Expected Value of Profits and the Cyclical Variability of Investment," unpublished paper, Harvard University, 1982.
- Clark, Peter K., "Investment in the 1970s: Theory, Performance, and Prediction," *Brookings Papers*, 1:1979, 73-113.
- Hall, Robert E. and Jorgenson, Dale W., "Tax Policy and Investment Behavior," *American Economic Review*, June 1967, 57, 391-414.
- Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Alan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl.

- 1976, 19-46.
- Musgrave, John C., "Fixed Capital Stock in the United States: Revised Estimates," *Survey of Current Business*, February 1981, 57-68.
- Sargent, Thomas J., "Estimation of Dynamic Labor Demand Schedules Under Rational Expectations," *Journal of Political Economy*, December 1978, 86, 1009-44.
- Summers, Lawrence H., "Taxation and Corporate Investment: A  $q$ -Theory Approach," *Brookings Papers*, 1:1981, 67-127.

# Welfare Aspects of Current U.S. Corporate Taxation

By ALAN J. AUERBACH\*

The corporate income tax has provided a steadily declining fraction of federal revenues during the postwar era, accounting for 23.0 percent during fiscal year 1966, and 14.2, 13.9, and 10.2 percent in 1971, 1976, and 1981, respectively. However, for several reasons, aggregate tax collections are of limited use in measuring the distortionary impact of the corporate tax.

First, one must allow for cyclical and, potentially, long-run changes in the rate of profit to measure changes in average tax rates. More fundamental problems lie in attempts to relate such an average tax rate to the desired measure, the marginal tax rate. Assets of different age cohorts historically have faced different tax schedules because of a succession of nonretroactive tax law changes. Average tax rates on current income represent an amalgam of such tax rules. Furthermore, even with a constant tax schedule, the "taxable income" which serves as the base of the corporate tax diverges systematically from economic income, because of the investment tax credit and the acceleration of depreciation allowances relative to actual economic depreciation. An asset typically receives deductions and credits that are more than sufficient to shelter all income associated with it in the years immediately after purchase, but faces a tax base exceeding economic income in more distant years, after all depreciation allowances have been used up. This implies that taxes will vary as a fraction of income depending on the age structure of assets. Fast growing firms will face low and potentially negative average tax rates on *current* income, though facing the same tax schedule as stagnant firms with positive current tax payments. (This ignores the further complications that arise from the asymmetric treatment accorded gains and

losses, which I discuss below.) Finally, there may be serious distortions in the allocation of capital *within* the corporate sector that aggregate tax rates, average or marginal, cannot measure.

These difficulties have led to the general use and acceptance of the "effective tax rate" to measure the distortionary effects of the tax system. This rate is based on a comparison of before- and after-tax internal rates of return on investment projects; letting  $r_g$  and  $r_n$  be these two measures, we define the effective tax rate to be

$$(1) \quad \tau = (r_g - r_n) / r_g.$$

If  $r_n$  is the return net of corporate taxes, then  $\tau$  measures the effective corporate tax rate. Setting  $r_n$  equal to the return after all income taxes produces a measure of the full tax wedge introduced by corporate and personal taxation.

In the remainder of this paper, I refer to such measures in analyzing the current impact of the corporate tax: in the aggregate, among different assets, and with respect to the treatment of the losses. I begin with a review of the evolution of thought on the additional burden imposed by the existence of a corporate income tax on top of an existing individual income tax.

## I. The Corporate Tax: Double (Taxation) or Nothing?

Certainly the most influential article on the corporate tax was Arnold Harberger's, which showed how an extra tax placed on capital income in one sector of a two-sector model would lead to shifts in production and prices. Harberger's view of the corporate tax as an extra tax stemmed from the implicit assumption that all individual capital income was taxed at a single rate. An obvious problem with this assumption is that not all equity

\*Yale University and The National Bureau of Economic Research.

income faces ordinary income taxation; capital gains are taxed at lower rates. More serious is the fact that corporations may use debt to finance their projects. Since interest payments are tax deductible, debt-financed corporate source income is taxable only to the ultimate recipient, and "double taxation" disappears. Thus, corporations have the *option* of being taxed like unincorporated businesses. If they don't choose this option, then the alternative must be preferable. For reasonable parameters, Joseph Stiglitz showed that corporations ought to use only debt for marginal finance, with the corporate tax serving as a tax only on intramarginal rents.

Still another reason why corporate source income may be taxed less heavily than first appeared to be the case relates to the way in which corporations finance investments when they do choose to use equity funds. The typical view of the burden of dividend taxation stems from the implicit assumption that new share issues provide the funds necessary for new investment projects. More realistic is the assumption that equity funds come from the additional retention of earnings. Not only does this dominate new issues (because of the income taxes that stockholders avoid), but it is also empirically the more common way funds are raised.

This fact has several implications. The most important in the current context (discussed by Mervyn King, David Bradford, and in some of my own earlier work, 1979a,b) is that dividend tax exerts *no* marginal effect on corporate behavior as long as retentions serve as the source of equity funds. It provides precisely the same treatment of retained earnings that households would receive under a personal consumption tax and currently receive through Individual Retirement Accounts: a deduction (reduction in dividend taxes) for savings when it occurs, and a tax on future dissaving.

Taking the two preceding arguments together, one obtains a theory suggesting that corporate income is taxed essentially once, regardless of the form it takes; debt is taxed at the personal level, equity at the corporate level (ignoring the relatively lightly taxed individual capital gains).

Having moved from double to single taxation by an examination of financial policy, we can continue to erode the tax base by looking at the real side. The results in this section so far presume a corporate *income* tax: a tax levied on a corporation's economic income. No such tax exists in the United States. Since 1954, there have been numerous changes in the law that have shortened the period over which the cost of depreciable assets may be deducted, and have permitted more generous patterns of allowances over given tax lifetimes. The investment tax credit (introduced in the year of Harberger's article, 1962) also provides a substantial reduction in the tax burden. If the combination of depreciation deductions and the investment tax credit offers the same tax savings as would allowances based on economic depreciation, assets face a tax burden equal to that an income tax would impose. (Indeed, this equivalence was what motivated the proposal by Dale Jorgenson and me to replace then current credits and depreciation schedules with a first-year write-off equal to the present value of economic depreciation, indexed for inflation.) At the other extreme (nowadays less extreme), a combination of credits and deductions producing tax savings equal to those obtained under immediate write-off, or "expensing" produce the same effect at the margin as abolition of the corporate tax (assuming equity finance is used): a zero effective tax rate. This is because, under expensing, government bears the same fraction of costs and receipts associated with investments. Just as with the tax on dividends, this is a case of "consumption-tax" treatment of capital income, with the same outcome concerning the effective marginal rate.

The Economic Recovery Tax Act of 1981 (ERTA) and the Tax Equity and Fiscal Responsibility Act of 1982 (TEFRA) have introduced important changes in the nature of depreciation allowances that assets receive. Under ERTA, depreciable assets were allocated essentially to three classes distinguished by tax lifetime, all very short relative to past practice: three and five years for equipment and fifteen years for structures. Under TEFRA, additional acceleration due

to occur in 1985 and 1986 was cancelled and a 50 percent basis adjustment for investment tax credits was introduced (i.e., depreciation allowance may now be taken only on 95 percent of the price of an asset qualifying for a 10 percent investment tax credit).

Many authors have calculated effective tax rates for particular assets and industries under ERTA, with the general finding that effective corporate tax rates for investment in equipment were negative, assuming equity finance (i.e., no additional tax savings from interest deductions) and using a variety of estimates about future rates of inflation (which matter because allowances still are not indexed for inflation). Even under TEFRA, this may still be the case. For example, Charles Hulten and James Robertson (1982) calculate that in the aggregate, the effective corporate tax rate on equity-financed equipment is now 3.5 percent for all nonresidential business and -0.6 percent for manufacturing, based on a projected 6 percent inflation rate, a 4 percent real return after corporate taxes and detailed estimates of rates of economic depreciation and capital stock composition. For nonresidential structures, the corresponding numbers are 36.3 and 39.6 percent, respectively. By any account, the aggregate tax rates based on these numbers for equipment and structures (15.8 and 11.7 percent, respectively) are far below the effective tax rates faced at any time during the pre-1981 postwar period. Indeed, if inflation were to disappear, the same assumptions suggest that the effective tax rate will be approximately zero in the aggregate. Thus, except for the effects of inflation, successive changes in the corporate tax have succeeded in removing the marginal tax on real corporate investment. This means that we must subtract another tax from each of the "single-tax" outcomes arrived at before. In the absence of inflation, personal equity income would not be taxed; corporate source debt income would be taxed at the personal level while facing a negative tax at the corporate level, with the net impact of taxation depending on the relative marginal tax rates on interest income at the corporate and personal levels.

## II. Misallocation under the Current Corporate Tax

While the preceding calculations give a sense of the overall burden and distortion facing corporate capital as a whole, one must also be concerned with the allocation of capital within the corporate sector. It is difficult to obtain a precise measure of the additional deadweight loss caused by the differential taxation of corporate capital. Particular movement away from uniform capital taxation need not make matters worse at all; some may actually improve economic efficiency, as we know from the theory of second best. However, a necessary first step to any welfare calculation is to determine which assets face high effective rates of tax relative to others. The calculations cited above clearly suggest that, for a *given* financial policy, nonresidential structures face a much heavier tax rate than equipment. But the implications of this result depend on whether there is a separation between real and financial decisions.

There are different models of the determination of financial structure. Some (as outlined by Merton Miller 1977) suggest that equilibrium will have the characteristic that individual firms will be indifferent between debt and equity finance, regardless of their debt-equity ratios. The personal tax advantage to equity will just offset the corporate tax advantage to debt. However, a more standard view is that not all of the corporate tax advantage to debt is offset, but that interior debt-equity ratios result because there are additional costs to leverage that increase as the amount of debt does. These costs may result either from an increased incentive for the firm to follow socially sub-optimal investment policies (as in Michael Jensen and William Meckling, 1976; and Stewart Myers, 1977) or the increased probability that the firm will make negative taxable income and not receive the full benefit of interest deductions (as in Harry DeAngelo and Ronald Masulis 1980). If these costs vary across assets, perhaps because of differences in the risk characteristics of associated returns, the tax advantage to debt might

be more fully available to some assets in a way that could exacerbate or mitigate apparent differences in effective tax rates. (Some of my own research, 1982, suggests that firms borrow proportionally *less* to invest in structures than equipment indicating that the tax structure may be more biased than the above calculations would indicate.)

### III. Should the Corporate Tax be Abolished?

At first blush, it seems hard to justify the continued existence of a tax that imposes a marginal rate close to zero (negative if one accounts for the additional tax savings produced by interest payments) but nevertheless distorts capital allocation through a schedule of widely varying effective tax rates for particular assets. One is tempted to paraphrase a familiar line from Shakespeare to characterize the current corporate tax as "full of sound and fury, collecting nothing." Yet its collections are not zero, even though small relative to the total federal budget. This is precisely because it is average and not marginal tax rates that matter when taxes are collected. Though the present value of taxes to be collected on *new* investments may be zero or even negative, assets in place will generate positive revenues that would be lost if the corporate tax were removed. This mechanism can be understood most easily in the context of a simpler tax system with no interest deductions and expensing of all investment. Under this scheme, new investments would generate zero tax liability in present value, but assets that received the expensing deduction in the past would pay positive taxes from the present on. Abolition of the corporate tax would simply provide a windfall to owners of these assets in much the same way as a reduction in dividend taxation would (and, presumably, did in 1981). It is misleading to conclude that such windfalls as might occur with a change in tax regime are "only transitory," since the revenue lost may require permanently higher tax rates on other income in the future. My simulation work with Laurence Kotlikoff (described in more detail in Kotlikoff's paper

in this session) suggest that this is an important issue empirically.

One further distinction between a system without a corporate tax and the current system is the asymmetry in the present treatment of gains and losses. This obviously would vanish if the corporate tax were repealed. The current system allows firms with current net operating losses to carry them "back" three years; that is, if the losses are exceeded by the sum of taxable income in the three years immediately preceding, the taxpayer may claim a refund retroactively against prior taxes, thereby receiving the same treatment as would occur under a system of direct refundability. If there are insufficient past profits to absorb current losses, the balance may be carried "forward," to be used in the future to offset taxable income. However, such "carry-forwards" suffer from two disadvantages: they expire after fifteen years and, perhaps more fundamentally, they do not accrue interest. Thus, the firm that carries losses forward can expect to receive an amount strictly less in present value than would be obtained through a direct loss offset.

This asymmetry in the tax law became more important with the passage of ERTA, since the acceleration of depreciation allowances put many firms, including very profitable ones simply making large additions to their capital stocks, at risk of suffering net operating losses. The tax law contained a provision, referred to as "safe-harbor leasing," that was intended to alleviate this problem, although TEFRA scaled back safe-harbor leasing through 1983 and repealed it thereafter.

### IV. Leasing and Treatment of Losses

The safe harbor leasing provisions of the Economic Recovery Tax Act relaxed certain restrictions on leasing and made it easier for one firm to "sell" its investment tax credits and depreciation deductions to another firm for a single initial cash payment plus future tax considerations. If it were possible for a firm to sell *all* of its losses in this way, then a system with full offset would result, provided

that there were enough taxable income in the economy to cover these losses and to provide for a competitive market in the purchase of such losses. However, the safe-harbor provisions and, indeed, leasing in general, can only be used to transfer losses to the extent that they stem from credits and deductions on new capital. Firms cannot dispose of accumulations of losses carried forward from the past or current losses over and above those generated by investment credits and depreciation deductions. This has led to a situation in which some firms have such a large overhang of tax losses to carry forward that they are essentially tax exempt, yet they can and do sell their credits and deductions. For these firms, safe-harbor leasing provides an initial cash subsidy to initial purchase of investment goods, the income of which will not be taxed. The resulting negative effective tax rates on such investments have been cited in criticism of leasing, but it is not clear why this outcome is undesirable. Because of the negative tax rates introduced by ACRS, such transfers to "tax-exempt" enterprises might still fall short of providing an effective tax rate as negative as those enjoyed by *taxable* companies. My analysis with Alvin Warren suggests that it is complicated to determine whether safe-harbor leasing transfers "too much" or "too little," in this sense, but that the leasing mechanism is ill-suited to accomplish such an objective.

Ultimately, however, one must ask why, if such equalization of effective tax rates is desired, it is not simply accomplished via a loss offset. One can construct models in which such asymmetry might be optimal, but this would argue against transferability through leasing as well as direct refundability. For example, suppose some firms are poorly managed in that they accept a lower rate of return in their project selection than their owners would wish; these firms overinvest relative to efficient levels. In such a situation, the need to carry losses forward could act as a corrective tax, falling most heavily on those firms likely to have losses frequently, that is, those that overinvest. (One would need to show here that this effect would outweigh the added incentive such

firms have to overinvest *after* losses occur, to use up accumulated losses before they lose value.) However, if such a structure is preferable to a system with full loss offset, there seems little case for the imperfect substitute for such an offset provided by leasing.

## REFERENCES

- Auerbach, Alan J., (1979a) "Share Valuation and Corporate Equity Policy," *Journal of Public Economics*, June 1979, 11, 291-305.
- \_\_\_\_\_, (1979b) "Wealth Maximization and the Cost of Capital," *Quarterly Journal of Economics*, August 1979, 93, 434-46.
- \_\_\_\_\_, "Real Determinants or Corporate Leverage," mimeo., 1982.
- \_\_\_\_\_, and Jorgenson, Dale W., "Inflation-Proof Depreciation of Assets," *Harvard Business Review*, September/October 1980, 58, 113-18.
- \_\_\_\_\_, and Kotlikoff, Laurence J., "Investment versus Savings Incentives: The Size of the Bang for the Buck and the Potential for Self-Financing Business Tax Cuts," Working Paper No. 1027, National Bureau of Economic Research, November 1982.
- Bradford, David, "The Incidence and Allocation Effects of a Tax on Corporate Distributions," *Journal of Public Economics*, April 1981, 15, 1-22.
- DeAngelo, Harry and Masulis, Ronald, "Optimal Capital Structure Under Corporate and Personal Taxation," *Journal of Financial Economics*, March 1980, 8, 3-81.
- Harberger, Arnold C., "The Incidence of the Corporation Income Tax," *Journal of Political Economy*, June 1962, 70, 215-40.
- Hulten, Charles R. and Robertson, James W., "Corporate Tax Policy and Economic Growth: An Analysis of the 1981 and 1982 Tax Acts," Discussion Paper, Urban Institute, December 1982.
- Jensen, Michael and Meckling, William, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics*, October 1976, 3, 305-60.
- King, Mervyn, "Taxation and the Cost of Capital," *Review of Economic Studies*,

January 1974, 41, 21–35.

Miller, Merton, "Debt and Taxes," *Journal of Finance*, May 1977, 32, 261–75.

Myers, Stewart, "Determinants of Corporate Borrowing," *Journal of Public Economy*, November 1977, 5, 147–75.

Stiglitz, Joseph E., "Taxation, Corporate

Financial Policy and the Cost of Capital," *Journal of Public Economy*, February 1973, 2, 1–34.

Warren, Alvin, C. and Auerbach, Alan J., "Transferability of the Tax Incentives and the Fiction of Safe-Harbor Leasing," *Harvard Law Review*, June 1982, 95, 1752–786.

# National Savings and Economic Policy: The Efficacy of Investment vs. Savings Incentives

By LAURENCE J. KOTLIKOFF\*

Investment incentives, as distinct from savings incentives, treat newly produced capital more favorably than existing capital. In the United States, accelerated depreciation allowances and the investment tax credit (*ITC*) are the most important examples of investment incentives. In the case of the *ITC*, the tax code restricts repeated use of the credit, regardless of increases over time in the size of the credit. Increases in the *ITC* discriminate, therefore, against existing capital that received a lower credit in the past. Enhanced acceleration of depreciation allowances is also injurious to old capital. The tax law requires the sale of old capital and the payment of recapture taxes prior to depreciating old capital under a new, more generous depreciation schedule.

Since equally productive units of new and old capital must sell for the same price, tax provisions favoring new capital imply a lower price for existing capital. The new investment incentives included in the 1981 Economic Recovery Tax Act (ERTA) provide an illustration of the potential capital losses involved. Alan Auerbach and I (1982) estimate that the 1981 Act reduced the value of existing plant and equipment by roughly \$290 billion, assuming no turnover of old capital. Under the assumption that all old capital was resold if the gain in investment incentives exceeded the recapture tax, the capital loss (inclusive of the recapture tax) equaled roughly \$230 billion. For U.S. wealth holders, then, the 1981 Tax Act imposed an implicit tax, in the form of a capital loss, ranging from \$230 to \$290 billion. These estimates assume zero marginal costs of adjusting the economy's capital stock. The assumption of substantial adjustment costs reduces these values by roughly one quarter.

Since the bulk of U.S. wealth is held by older age groups, the 1981 legislation transferred resources away from this segment of the population. The recipients of this transfer were the current young and future generations, since these generations can now acquire title to productive capital at a reduced price.

Investment incentives, like government surpluses, transfer resources from older to younger generations. It is this redistribution through the revaluation of old relative to new capital that distinguishes investment from savings incentives. While both investment and savings incentives alter household marginal incentives to accumulate more capital, the income effects associated with intergenerational transfers potentially make investment incentives much more powerful devices in promoting capital formation. These effects are particularly strong in nonaltruistic life cycle economies in which the old, with fewer remaining years to consume, have greater marginal propensities to consume out of lifetime resources than do the young.

The extent to which intergenerational transfers, whether associated with explicit budget deficits, implicit deficits, such as unfunded Social Security, or tax-induced recapitalizations, alter wealth accumulation remains a matter of considerable controversy (Martin Feldstein, 1974; Robert Barro, 1974). Lawrence Summers and I (1981) suggest a predominate role for altruistic private intergenerational transfers in explaining the current stock of U.S. wealth; longitudinal age-earnings and age-consumption profiles are far from consistent with predictions of the strict, non-altruistic life cycle model (Franco Modigliani and Richard Brumberg, 1954). This finding does not, however, preclude the possibility that the majority of households conform to the selfish life cycle model. The majority of households could have such pref-

\*Yale University and The National Bureau of Economic Research.

erences, but simply have very little "hump" savings. It may well be that we live in a mixed society consisting of a minority of quite wealthy, altruistic households, and a majority of rather poor, life cycle households. While life cycle households may be responsible for little if any of the current stock of wealth, their response to new government policies, in particular, intergenerational transfers, could well dictate the economy's short-run saving behavior.

Given our state of ignorance concerning the distribution of intertemporal preferences, exploring the implications of investment incentives in a strictly life cycle model can be justified as providing an upper-bound estimate on the potential response to investment policy. The life cycle model is also a convenient framework for expositing the hidden surpluses imbedded in "business tax cuts." Given the nation's current pre-occupation with projected budget deficits, understanding that the federal government in 1981 effectively collected \$230 to \$290 billion more in revenue than it reported seems quite important.

A verbal explanation of this last statement provides an intuitive introduction to the simple model presented below. In 1981, the official U.S. federal deficit equaled \$58 billion. During that year the government implicitly imposed a \$230 to \$290 billion tax on existing capital by enacting ERTA. Had the government explicitly imposed this wealth tax, reported government receipts would have risen by perhaps, \$260 billion, and the government would have reported a \$202 billion surplus for 1981. Levying an explicit rather than implicit \$260 billion tax on wealth would have left wealth holders no worse off and should have been a matter of indifference to them.

Younger generations could also have been made equally well off had the implicit tax been made explicit. Under the implicit tax, the young pay lower effective capital income taxes on their investment in capital because its acquisition price is subsidized if they buy new capital, or reduced if they buy old capital. Under the explicit tax scenario, the government retains the wealth tax revenues as a surplus, and the young purchase new

and old capital at its previous value. However, if the government uses annual interest income earned on the wealth tax to reduce capital income taxes, the young end up facing the same effective lifetime capital income taxation as in the implicit tax case.

Unfortunately, conventional government accounting procedures fail to record such hidden surpluses or deficits, while making great moment of "official" budget deficits. The reality of postwar fiscal history is that implicit deficits and surpluses associated with unfunded government retirement programs and changes in investment incentives greatly swamp official reported deficits when measured by their potential affect on capital formation and real interest rates (1982 *Economic Report of the President*, Appendix to ch. 4; Auerbach's and my two forthcoming papers).

### I. Investment Incentives

A simple two-period life cycle model of economic growth provides a convenient framework for examining the underlying nature of investment incentives. Consider such an economy with a tax  $\tau_{y,t}$  on business profits, and an investment incentive, which, for simplicity, is an expensing rate for new capital of  $e_t$ . The subscript  $t$  denotes the period in which the two instruments are applied. To simplify the analysis further, assume individuals work only when young and that the depreciation rate, the rate of population growth, and the rate of technological change are zero.

Equations (1) and (2) characterize the economy's process of capital formation:

$$(1) \quad K_t = (W_{y,t-1} - C_{y,t-1})/q_{t-1},$$

$$(2) \quad C_{0,t} = q_{t-1}K_t(1+r_t).$$

In (1),  $W_{y,t-1} - C_{y,t-1}$  is the saving of the young in period  $t-1$ , their after-tax wages in period  $t-1$  less their consumption in period  $t-1$ . The net price of a unit of capital in period  $t-1$  is given by  $q_{t-1}$ . Dividing the financial saving of the young by  $q_{t-1}$  determines their purchase of physical capital (assuming there is no government debt or

other assets in the economy). The physical capital acquired by the young at the end of period  $t-1$  equals the economy's capital stock at the beginning of period  $t$ ,  $K_t$ ; that is, the old generation in period  $t$ , those young in  $t-1$ , hold claims to all the economy's capital, since the young in period  $t$  have no beginning of period assets.

For the old in period  $t$ , consumption,  $C_{0,t}$ , equals the return of principal,  $q_{t-1}K_t$ , plus the after-tax return on the investment,  $q_{t-1}K_t r_t$ . The after-tax return,  $r_t$ , includes capital gains and losses:

$$(3) \quad r_t = \frac{F_{K,t}(1 - \tau_{r,t}) + q_t - q_{t-1}}{q_{t-1}}$$

where  $F_{K,t}(1 - \tau_{r,t})$  equals marginal after-tax profits per unit of capital. In combination, (2) and (3) imply

$$(4) \quad C_{0,t} = q_t K_t + K_t F_{K,t}(1 - \tau_{r,t}).$$

This new expression is also intuitive: the consumption of the old in period  $t$  (the young of period  $t-1$ ) equals after-tax business profits plus the value of the sale of their capital at the prevailing asset price  $q_t$ .

Equation (5) expresses  $q_t$ , the net price of purchasing a unit of capital, in terms of  $\tau_{r,t} e_t$ :

$$(5) \quad q_t = 1 - \tau_{r,t} e_t.$$

For new capital, the net acquisition cost is 1, the price of new capital, less the tax rebate from expensing  $\tau_{r,t} e_t$ . Equation (5) also determines the price of old capital. Since old capital and new capital are perfect substitutes in production their net acquisition costs must be identical in equilibrium; hence, old capital sells for  $\tau_{r,t} e_t$  less than new capital. Equations (1), (4), and (5) may now be combined to indicate the lifetime budget constraint of the young in period  $t-1$ ,

$$(6) \quad C_{y,t-1} + C_{0,t} \frac{(1 - \tau_{r,t-1} e_{t-1})}{(1 - \tau_{r,t} e_t) + F_{K,t}(1 - \tau_{r,t})} = W_{y,t-1}$$

and the old in period  $t-1$ :

$$(7) \quad C_{0,t-1} = K_{t-1}(1 - \tau_{r,t-1} e_{t-1}) + K_{t-1} F_{K,t-1}(1 - \tau_{r,t-1}).$$

The essential feature of investment incentives can be illustrated most simply by assuming permanent capital income taxation at rate  $\tau_r$ , zero expensing prior to period  $t-1$ , and a permanent move to 100 percent expensing starting at time  $t-1$ . Under these assumptions, all tax terms drop out of equation (6); the young of period  $t-1$  and all future generations face zero effective taxation over their lifetimes. While the young and future generations nominally pay business profits taxes in their old age, the reduced cost of purchasing capital when they are young exactly offsets the present value cost of this taxation. Stated differently, new generations starting in year  $t-1$  are subsidized when young to purchase capital and taxed when old on its return. The subsidy and tax cancel in present value and the young face no net taxation on their capital investments.

While this new tax structure effectively exempts the young of period  $t-1$  and all future generations from paying any taxes over their lifetime, elderly individuals at time  $t-1$  suffer a capital loss on their assets equal to  $K_{t-1} \tau_r$ . According to (7), the consumption of the elderly falls by this amount; the  $K_{t-1} \tau_r$  capital loss constitutes a one-time wealth tax on the old of period  $t-1$ . Considering the tax treatment of the young and old together, this new tax system is equivalent to the government's collecting  $K_{t-1} (1 + F_{K,t-1}) \tau_r$  in taxes from the old in period  $t-1$  and abolishing taxation thereafter. If it so choose, the government could adjust its consumption expenditures each year to just equal interest earned on the government surplus. In this case, the government, according to standard accounting procedures, would report a surplus in period  $t-1$ , and a balanced budget thereafter. The alternative method of subsidizing the young to purchase capital and taxing them in their old age on its return is effectively equivalent to the government

using each young generation as its savings account. In this case of an implicit wealth tax, the government reports budget balance in period  $t - 1$  as well as all future periods.

This model is also useful for describing savings incentives. By definition, such incentives leave the relative price of old to new capital unchanged. A simple example of a savings incentive is a switch from capital income to wage income taxation in a world in which the expensing rate is zero. Obviously there is a myriad of other savings incentive policies which leave this relative price unchanged. For example, reductions in capital income taxes could be temporary or permanent, and these reductions could be associated with short-term deficits, increases in alternative taxes, reductions in government consumption, etc.

#### *A. Investment vs. Savings Incentives — Illustrative Simulations*

Given the range of concomitant government policies that could and would be adjusted in response to either investment or savings incentives, unconditional statements such as "investment incentives stimulate more capital formation than savings incentives" are meaningless. Analysis of investment and savings incentives for explicitly specified policies of adjusting to the associated revenue changes do, however, permit meaningful conditional comparisons of these policies.

The Auerbach-Kotlikoff dynamic simulation model described in our 1983a paper and our paper with J. Skinner, provides a framework for analyzing conditional policy experiments. The model computes the perfect foresight equilibrium path of neoclassical economies during transitions between steady states. The life cycle version of the model used to study investment and savings incentives incorporates the following government policy instruments: expensing, capital income, consumption, and wage taxation, and the choice of a deficit policy. The household sector of the economy consists of fifty-five overlapping generations making life cycle consumption and labor supply decisions based on an intertemporally separable CES

utility function. The simulations discussed here assume a 1 percent population growth rate, a static elasticity of substitution of .8 and an intertemporal elasticity of substitution of .25. The production function is Cobb-Douglas with capital's income share equal to .25. In the economy's initial steady state, government consumption per capita is financed by a uniform 30 percent proportional income tax, and there is no initial government debt.

The following is one example of a savings incentive: capital income taxes are permanently cut to zero; the short fall in revenue is financed by deficits for the first five years; after year five, wage tax rates are increased to preclude future growth in per capita debt as well as to finance the constant stream of precapital government consumption. As a result of this policy, the aggregate capital stock rises by almost 10 percent over the first five years and then declines. Eventually the capital stock declines some 6 percent below its initial level.

Consider a comparable investment incentive policy. The government switches immediately from zero to 100 percent expensing in the first year and finances this over a five-year period by increases in the national debt. Thereafter, income taxes are raised to hold constant the per capita stock of debt as well as to finance the same constant time path of government consumption as in the savings incentive policy.

The difference in the impact on capital accumulation is striking. In the first five years, the capital stock grows slightly faster than with the savings incentive policy and thereafter it continues to grow. After twenty years, the percentage rise is some 40 percent and the eventual increase exceeds 50 percent.

The two policies also have markedly different welfare implications. The investment policy results in a long-run welfare gain (measured as a compensating variation in full-time resources in the initial steady state) of 6 percent; the long-run welfare loss under the savings policy is 9 percent. Under the savings policy, the wage tax rate rises in the long run to 48 percent; in contrast, the increase in the tax base under the investment policy actually permits a long-run reduction

in the rate of income taxation from 30 to 29 percent.

### *B. Investment Incentives and Strategies for Dealing with Deficits*

Various strategies have been offered to reduce short-run deficits associated with business tax cut policies. One solution to short-run revenue losses is a phase-in of investment incentives. This characterized the Economic Recovery Tax Act of 1981, which called for the acceleration of depreciation allowances to increase in 1981 and again in 1985 and 1986. The problem with policies of this kind is that they induce capital losses gradually over the phase-in period (consider equation (5) for  $e_t$  increasing through time). Investor's projections of these future capital losses discourages investment in the short run, defeating the entire purpose of the legislation. As an example, consider the effects of a five-year phase-in of expensing without deficits, with the expensing fraction rising linearly from .2 in the first transition year to 1 in the fifth. Though investment eventually expands under this policy, the short-run impact is to discourage investment. Investment stays essentially flat until the fourth year of the policy.

A more successful way of avoiding deficits recognizes that investment incentives can often be achieved by raising rather than lowering capital income taxes. As indicated by equation (5), given a positive rate of expensing, the increase in the statutory capital income tax rate increases the implicit wealth tax on existing capital. This reduces the consumption of wealth holders, permitting an expansion of national saving and investment. In addition, the extra revenue from the capital income tax allows the government to lower other taxes. Starting from an initial steady state with full expensing and a 30 percent income tax, raising the capital income tax rate to 50 percent permits an immediate drop in the wage tax rate to 26.5 percent (falling eventually to 21.6 percent) to maintain budget balance and an eventual increase in capital per person of 34.6 percent.

Finally, investment incentives may be self-financing in the long run, requiring no current or future increase in statutory tax rates to achieve a more capital intensive long-run steady state. Consider a policy of moving directly from zero to 50 percent fractional expensing, with the income tax held constant at 30 percent for twenty years; while this policy generates short-run deficits, the expansion of the income tax base over time raises revenue sufficient to retire this debt. Indeed, in the twentieth year, the debt-capital ratio is  $-.36$  percent. This surplus permits a slight decrease in the income tax rate thereafter (to avoid an expanding surplus), to 29.2 percent in the twenty-first year and 29.0 percent in the long run. The per capita capital stock increases by 25.9 percent in the long run.

While taxes on capital income and, eventually, labor income decline, existing capital owners do face increased implicit wealth taxation under "self-financing business tax cuts." Their welfare declines, thus distinguishing this policy from the "free lunch" promised by certain "supply-side theorists." Another reason that business tax cuts can be self-financing is that the economy has shifted to a more efficient tax structure, substituting lump sum taxes on initial wealth holders for distortionary income taxes on current and future generations. These efficiency gains, in addition to the implicit, but real transfers from the initial elderly, provide the economic resources to "cut taxes and raise revenues."

## **II. Summary**

The key difference between savings and investment incentives in closed economies is the applicability of these incentives to old as well as new capital. Investment incentives discriminate against old capital; savings incentives do not. This discrimination reduces the market value of old capital and, therefore, the economic resources of owners of the existing capital stock. The reduction in the resources and welfare of initial wealth holders under investment policies are similar, if not identical, to those arising from a one-time wealth tax.

In life cycle economies, the remaining resources of the elderly are held primarily in the form of nonhuman wealth. The wealth tax generated by investment incentives falls, therefore, most heavily on the elderly. Since the elderly, in life cycle models, have a greater marginal propensity to consume than young and future generations, this intergenerational redistribution of resources away from the elderly reduces current aggregate consumption. The reduction in the consumption of the elderly effectively finances the "crowding in" of investment, and explains the extra 'bang for the buck.'

#### REFERENCES

- Auerbach, Alan J. and Laurence J. Kotlikoff, (1983a) "National Savings, Economic Welfare, and the Structure of Taxation," in M. Feldstein, ed., *Behavioral Simulation Methods in Tax Policy Analysis*, Chicago: University of Chicago Press, forthcoming 1983.
- \_\_\_\_\_, and \_\_\_\_\_, (1983b) "An Examination of Empirical Tests of Social Security and Savings," in E. Helpman, et al., eds., *Social Policy Evaluation: An Economic Perspective*, New York: Academic Press, forthcoming 1983.
- \_\_\_\_\_, and \_\_\_\_\_, "Investment versus Savings Incentives: The Size of the Bang for the Buck and the Potential for Self-Financing Business Cuts," Working Paper No. 1027, National Bureau of Economic Research, November 1982.
- \_\_\_\_\_, \_\_\_\_\_, and Skinner J., "The Efficiency Gains from Dynamic Tax Reform," *International Economic Review*, forthcoming 1983.
- Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, 82, 1095-117.
- Feldstein, Martin S., "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *Journal of Political Economy*, September/October, 1974, 82, 905-26.
- Kotlikoff, Laurence J., and Summers, Lawrence, "The Role of Intergenerational Transfers in Aggregate Capital Formation," *Journal of Political Economy*, August 1981, 89, 706-32.
- Modigliani, Franco and Brumberg, Richard, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in K. Kurihara, ed., *Post Keynesian Economics*, New Brunswick: Rutgers University Press, 1954.
- U.S. Council of Economic Advisers, *Economic Report of the President*, Washington: USGPO, 1982.

## RECENT STRUCTURAL CHANGE IN THE CAPITAL MARKETS

### The Process of Financial Innovation

By WILLIAM L. SILBER\*

We often think of technological changes, such as the telegraph or xerography, as *de novo* events. That is no doubt true in some cases, but most evidence suggests that innovative activity responds to economic forces. In the financial sector, casual observation credits most new financial products to economic incentives, but there is a paucity of formal empirical evidence on the subject. For example, many agree that high interest rates and regulation stimulated new financial products, such as money market funds and NOW accounts, but we need both theoretical models and empirical observation to document the forces at work.

The objective of this paper is to survey recent financial innovations and to provide a framework for understanding why they arose. Models are designed to explain the process of financial innovation and are used to identify the variables that underlie new financial products and practices.

#### I. Theories of Financial Innovation

Financial innovation is often viewed as a by-product of regulation. The argument is that most financial innovations try to circumvent regulatory constraints. There is little doubt that regulations play an important role, but that perspective is too narrow to explain the entire process. A more general model (see my 1975 article) has been developed emphasizing the microeconomic framework of financial innovation. I summarize

that approach because it helps pinpoint the sources of innovative pressure.

The main hypothesis (1975, pp. 64ff.) is quite straightforward: new financial instruments or practices are innovated to lessen the financial constraints imposed on firms. In fact, a simple linear programming model helps to articulate the process. Firms maximize utility subject to a number of constraints. These constraints are imposed both externally and internally. Among the most prominent external constraints are government regulations. But the marketplace also constrains the firm's optimization problem. For example, if market power exists, then the firm sets prices or yields and accepts whatever volume of funds are offered; alternatively if the firm is a price taker, it can buy precisely the quantity of funds that maximizes utility. Viewed more generally, the marketplace defines the parameters of demand and supply for different financial products and simultaneously identifies the policy tools available to the firm.

Constraints are often set internally by the firm, and these also influence the optimization problem. For example, a firm may establish a target rate of growth for assets. Self-imposed liquidity constraints are also frequently found within financial firms. Commercial banks, for example, usually manage discretionary funds so that all potential sources of funds are tapped on a regular basis (see my 1978 study, p. 15).

In the normal course of events, a firm maximizes its objective function subject to existing constraints. This simply says that a firm will sell securities or accept deposits and invest the proceeds, all within the framework of existing parameters and constraints. New sources and uses of funds are innovated when exogenous changes in the environment stimulate the search for new policy tools. As

\*Professor of economics and finance, Graduate School of Business, New York University, and research associate, NBER. I wish to thank Ben Friedman, Ken Garbade, Dwight Jaffee, and Robert Kavesh for helpful comments. Financial support for this paper was received from NSF grant No. SES-8103156 and from the Bankers Trust Co.

long as the development of new financial products is costly, normal financial decision making will be distinguished from new product innovation.

The programming framework suggests that the stimulus to innovation can be interpreted as an increase in the cost of adhering to existing constraints. For example, rising yields on assets in the objective function may raise the cost (shadow price) of the constrained volume of deposit funds. Alternatively, an increase in interest rate volatility may reduce funds from conventional fixed rate debt instruments, implying a higher shadow price for the proceeds of conventional securities flotations. In both cases, the rising costs of adhering to existing financing constraints stimulates the search for new financial products.

This constraint-induced innovation hypothesis acquires a time dimension once search and development costs are integrated with rising shadow prices. Although financial innovation is not considered as costly as new technology, there is considerable expense in designing new contracts, running a new secondary market or installing requisite computer equipment. These costs suggest that only a sustained increase in shadow prices over time will stimulate new product innovation. In addition, lower costs of developing new products will generate more innovations.

Before turning to the empirical evidence, two extensions of this view of financial innovation are useful. One explicit addendum is Richard Sylla (1982) and relates specifically to collective action and monetary innovation. The second observation is implicit in Albert Wojnilower's (1980) treatment of credit crunches and deals with the nature of new financial products.

Sylla's basic point is that new monetary standards (fiat currency, national bank notes) were introduced when crises in the monetary system forced the government to overhaul the payments mechanism. Sylla (especially fn. 6) follows my study and identifies the innovation of new monetary systems with the rising cost of adhering to existing constraints, where the rising costs are felt by individuals and firms simultaneously. Con-

sidering the congenital difficulty in both public and private collective action, it is not surprising that crises are needed to induce separate interest groups to innovate a product jointly.

Wojnilower's observations on new financial products stem from his interest in how financial responses to credit crunches set the stage for future cycles of recession and inflation (p. 227). From my standpoint, his most interesting observation (p. 295) is that floating rate debt transforms a lender's objective into simply maximizing total assets, rather than the normal constrained optimization. Wojnilower's perspective fits quite well within my framework, but emphasizes that new financial products are designed to *sustain* financing flexibility for the firm.

## II. Empirical Evidence on Sources of Innovation

The only formal empirical investigation of financial innovation appears in Moshe Ben-Horim's and my 1977 article. To test the constraint-induced model of innovation, that study specified a linear programming model of large money market banks designed to calculate shadow prices of deposits and capital between 1952 and 1970. The results were quite encouraging. Deposit shadow prices rose significantly in the years prior to 1961 and again before 1969. Both years experienced major financial innovations: in 1961, the negotiable certificate of deposit was born; and in 1969, bank-related commercial paper and loan repurchases were popularized. The shadow prices for capital jumped significantly between 1962 and 1964. In 1963, subordinated debentures were first introduced as part of bank capital, and in 1965, those debentures were issued in substantial volume. Thus the empirical results confirm that rising costs of adhering to constraints stimulate financial innovation.

Table 1 records new financial products and practices between 1970 and 1982. This provides a new data base on which to test the constraint-induced model of the innovative process. Most of the innovations listed in the table were cited in articles (available on request) from the *Wall Street Journal*, *Business Week*, *American Banker*, *Institu-*

TABLE 1—FINANCIAL INNOVATIONS: 1970–82

Types	Exogenous Causes <sup>a</sup>					
	1			2	3	4
	(a)	(b)	(c)			
A. Cash Management						
1. Money Market Mutual Funds	✓					
2. Cash Management/Sweep Accounts	✓				✓	
3. Money Market Certificates	✓					✓
4. Debit Card	✓				✓	
5. NOW Accounts	✓					
6. ATS Accounts	✓				✓	
7. Point of Sale Terminals					✓	
8. Automated Clearing Houses					✓	
9. CHIPS (Same Day Settlement)					✓	
10. Automated Teller Machines					✓	
B. Investment Contracts						
(i) Primary Market						
1. Floating Rate Notes				✓		
2. Deep Discount (Zero Coupon) Bonds	✓		✓	✓		
3. Stripped Bonds	✓		✓	✓		
4. Bonds with Put Options or Warrants	✓			✓		
5. Floating Prime Rate Loans				✓		
6. Variable Rate Mortgages				✓		
7. Commodity Linked (Silver) Bonds				✓		
8. Eurocurrency Bonds	✓					✓
9. Interest Rate Futures				✓		
0. Foreign Currency Futures						✓
1. Cash Settlement (Stock Index) Futures						✓
2. Options on Futures				✓		✓
3. Pass-Through Securities						✓
(ii) Consumer-Type						
1. Universal Life Insurance				✓		
2. Variable Life Policies		✓				
3. IRA/Keogh Accounts			✓			✓
4. Municipal Bonds Funds			✓			✓
5. All-Saver Certificates	✓					✓
6. Equity Access Account		✓		✓		
C. Market Structures						
1. Exchange-Traded Options						✓
2. Direct Public Sale of Securities						
Green Mountain Power Co.				✓		
Shelf Registration				✓		✓
3. Electronic Trading						
NASDAQ					✓	
GARBAN					✓	
4. Discount Brokerage						✓
5. Interstate Depository Institutions					✓	✓
D. Institutional Organization						
1. Investment Bankers/Commodity Dealers	✓			✓		
Salomon/Phibro, Goldman						
Sachs/J. Aron, DLJ/ACLI						
2. Brokers/General Finance						✓
Shearson/Amex, Bache/						
Prudential, Schwab/Bank of America						
3. Thrifts with Commercial Banks	✓			✓		✓
4. Financial Centers (Sears Roebuck)						✓

<sup>a</sup>Column headings: 1. Inflation: (a) Level of Interest Rates, (b) General Price Level, (c) Tax Effects; 2. Volatility of Interest Rates; 3. Technology; 4. Legislative Initiative; 5. Internationalization; 6. Other.

*tional Investor, Bank Marketing, Best's Review*, and other trade publications. My use of trade journals for compiling a list of innovations follows Edwin Mansfield (1968) who prefers that source to patent data (even when the latter are available).

The 38 individual entries in Table 1 are new financial products or practices that emerged during the 1970–82 period. There is little doubt that any single entry could be expanded into additional specific innovations. For example, interest rate futures (entry B(i),9) could be divided into numerous individual new futures contracts. Although such distinctions are quite important in certain contexts (see my 1981 article), my broad objectives required that these details be left to more specific analyses.

I cannot replicate the Ben-Horim-Silber experiments for the innovations recorded in Table 1. Considerable additional effort is needed to construct a time-series of shadow prices for each innovation. Instead, I take a less formal approach to evaluating the constraint-induced explanation of financial innovation. In particular, each of the columns in the table represents an exogenous force that influenced financial constraints during the 1970–82 period. I would like to see how many new financial products were stimulated by these forces and how well the constraint-induced model of innovation tells the story. Needless to say, this evidence on the innovative process is best viewed heuristically.

Each of the potential causes of innovation requires some elaboration. The first two items, inflation and volatility of interest rates, are self-explanatory, although I identify interest rate and tax burden effects of inflation in addition to the general increase in prices. Technology refers primarily to information processing and data transmission. Legislative initiatives are identified with agency regulations or congressional legislation. Internationalization includes the consequences of the expansion in foreign trade, floating exchange rates, and OPEC. Although I have labelled each of these items exogenous, some feedback effects will be noted as I discuss the individual new products and practices.

Cash management is the first broad category of innovation listed in the table. More particularly, the first six entries (A.1–A.6) are varieties of cash management accounts. The high level of interest rates is listed as a primary cause of each item. For three of these innovations, sweep accounts, debit cards, and ATS accounts, I also note the role of inexpensive computer technology (this might also be true of money market mutual funds). The money market certificate was also a by-product of interest rates levels, with an assist from favorable regulatory action (although this probably should not be labelled a legislative *initiative*).

All of these innovations were aimed at loosening constrained access to sources of funds created by rising interest rates. Even money market mutual funds might be viewed within this context because the industry responded to lagging equity fund sales caused, in part, by rising rates of interest. Note that the cost side of circumventing constraints enters via the role of inexpensive computer technology.

Items A.7 through A.10 are new products associated with the technology of the payments system. It is possible that interest rate levels played some role in stimulating the same day settlement procedure of CHIPS (Clearinghouse Interbank Payment System), but technology was the major force behind point of sale terminals, automated clearing houses and automated teller machines.

The next group of innovations is labelled (B) investment contracts. This category is divided into (i) primary market innovations and (ii) consumer-oriented products. While there are similarities in the subcategories, it is useful to consider them separately.

The first eight innovations under primary market investment contracts are modified debt instruments designed to help borrowers raise funds in an environment of high and volatile interest rates. In particular, 1) floating rate notes, 4) bonds with put options (bonds that can be sold back to the corporation at a fixed price), 5) the floating prime, 6) variable rate mortgages, and 7) commodity-linked bonds protect lenders from rising interest rates. Without such flexibility, bor-

rower access to funds would have been severely constrained. Similarly, 4) bonds with warrants (to buy additional debt from the corporation at a fixed yield), 2) deep discount bonds, and 3) stripped bonds (where the semiannual coupons and the final payment have been separated) also were designed because of interest rate volatility (in addition to offering certain tax benefits). In these cases, investors can lock in exceptionally high yields. Thus, the table indicates that both the volatility in yields and the high level of rates were important stimuli for these innovations. Eurocurrency bonds (item 8) are a product primarily of the increased internationalization of corporate balance sheets, but without high and volatile domestic interest rates, this market would probably not have grown significantly.

Four new products in futures markets are listed as items 9 through 12 under category B(i). Although these are not directly related to reducing the burden of financing constraints, they are complementary innovations that permit investors to hedge the risk of increased price volatility in financial markets. Note that both cash settlement (as opposed to physical delivery) futures contracts and options on futures required regulatory approval. But as with money market certificates, the initiative hardly came from the regulators.

Pass-through securities, listed as entry 13, is the first item in the table that emerged *primarily* through government initiative, via the GNMA pass-through program of the Department of HUD. Private pass-throughs subsequently joined them in the marketplace. The pass-through program was designed to increase the liquidity of mortgages and to direct the flow of credit to the mortgage market (see Rudolph Penner's and my 1973 article, and Deborah Black et al., 1981). It is unclear whether this is a constraint-induced innovation initiated by government collective action.

Under consumer-type innovations of subcategory B(ii), three items were initiated by legislation: 3) IRA/Keogh accounts; 4) municipal bond funds; and 5) all-saver certificates. But even these legislative

initiatives are not truly exogenous. With IRA/Keogh accounts and municipal bond funds, the rising tax burden associated with inflation probably stimulated legislative relief. For the all-saver certificate, the main thrust was to alleviate thrift institution funding constraints.

The remaining three innovations in the consumer-type category fit easily into the constraint-induced framework. Universal life insurance (item 1) combines insurance with a money market fund so that insurance companies can compete with short-term investments. Variable life policies (entry 2) keep insurance coverage constant in real terms, permitting insurance companies to increase nominal cash inflows during inflationary periods. The equity access account (item 5) was designed by Merrill Lynch & Co. to permit consumers to overcome liquidity constraints by borrowing through a line of credit against the built-up equity in their homes.

Approximately 75 percent of the new products listed under categories A and B fit into the constraint-induced framework. The model does not do nearly as well in the next two categories: (C) market structures and (D) institutional organization. Each requires some elaboration.

Among the new market structures listed under C, only item (2), the direct public sale of securities, explicitly promotes flexibility in financing constraints. The Green Mountain Power Co. of Vermont began the process back in 1970 by selling debt directly to its customers on a continuous basis. Sears Roebuck and AT&T announced similar plans but never consummated the arrangements. A ruling by the SEC now permits firms to register securities for sale over a two-year period and then to sell them "off the shelf" as market conditions dictate. This flexibility in the timing of new issues is crucial when interest rates are volatile.

The other developments in new market structures have little to do with relaxing constraints. Exchange traded options (entry 1) was an experiment of the Chicago Board of Trade through the Chicago Board Options Exchange. Electronic trading (item 3) in over the counter stocks (NASDAQ) and in the

government securities market (GARBAN) were technological breakthroughs. The SEC's end to fixed commission rates on May 1, 1975 initiated the era of discount brokerage (item 4). Finally, de facto interstate depository institutions (entry 5) stem from technological developments and implicit regulatory approval.

Constraints play a greater role in category D, institutional organization. The increased level and volatility of interest rates together with the development of financial futures markets encouraged the mergers between commodities dealers and investment bankers (entry D.1), thereby promoting institutional flexibility. A similar argument might apply to the merger between thrifts and commercial banks in item 3. But there is no constraint-induced story behind the takeover of brokerage houses by other financial institutions in entry 2 or financial centers in item 4. These developments rely on an anticipated synergism that has yet to emerge.

### III. Conclusions

I can summarize the empirical evidence on innovation with three points. First, the only formal test of the constraint-induced model of financial innovation appears in Ben-Horim's and my article. The model successfully explained new bank products during the 1952-70 observation period. Second, the constraint model underlies about 60 percent of the financial innovations during the 1970-82 period. Although formal work estimating the relevant shadow prices is necessary, the constraint-induced story seems appropriate in at least half of the cases. Third, the most important forces underlying the remaining innovations are technology and legislative initiatives. Both factors sometimes operate through constraints, but it is a mistake to ignore their independent role in financial innovations.

I began this paper by comparing financial innovation with technological change. I noted that both types of innovative activity respond to economic incentives. I conclude by mentioning that both processes also improve economic welfare. Technological change ex-

pands physical output, thereby increasing the standard of living. Financial innovation operates in somewhat different dimensions, as emphasized by Benjamin Friedman (1982, pp. 53-56). In particular, innovations in financial institutions and practices have improved the ability to bear risk (futures markets), lowered transactions costs (automated teller machines), and circumvented outmoded regulations (money market mutual funds and Regulation Q). Thus, the process of financial innovation described here yields economic benefits that are no less real in a welfare sense than improvements in physical technology.

### REFERENCES

- Ben-Horim, Moshe and Silber, William, "Financial Innovation: A Linear Programming Approach," *Journal of Banking and Finance*, 1977, 1, 277-96.
- Black, Deborah, Kenneth Garbade and William Silber, "The Impact of the GNMA Pass-Through Program on FHA Mortgage Costs," *Journal of Finance*, May 1981, 36, 457-69.
- Friedman, Benjamin, "Postwar Changes in the American Financial Markets" in Martin Feldstein, ed., *The American Economy in Transition*, Chicago: University of Chicago Press, 1980.
- Mansfield, Edwin, *The Economics of Technical Change*, New York: W. W. Norton & Co., 1968.
- Penner, Rudolph and William Silber, "The Interaction between Federal Credit Programs and the Impact on the Allocation of Credit," *American Economic Review*, December 1973, 63, 838-52.
- Silber, William, "Towards a Theory of Financial Innovation" in his *Financial Innovation*, Lexington: D. C. Heath & Co., 1975.
- \_\_\_\_\_, *Commercial Bank Liability Management*, Chicago: Association of Reserve City Bankers, 1978.
- \_\_\_\_\_, "Innovation, Competition, and New Contract Design in Futures Markets," *Journal of Futures Markets*, 1981, 2,

125-55.

Sylla, Richard, "Monetary Innovations and Crises in American Economic History," in P. Wachtel, ed., *Crises in the Economic and Financial Structure*, Lexington: D. C.

Heath & Co., 1982.

Wojnilower, Albert, "The Central Role of Credit Crunches in Recent Financial History," *Brookings Papers on Economic Activity*, 2:1980, 277-326.

# Policy Implications of Structural Changes in Financial Markets

By EDWARD J. KANE\*

Economic decisions are goal-seeking choices made under acknowledged constraints. For financial institutions, models of managerial decision making feature environmental constraints of three broad types: technological; market; and regulatory. In principle, a shift in any constraint triggers a reactive reoptimization of an institution's product line, organizational structure, production processes, and demand for factor services.

Monetary policy and deposit-institution regulation promote three major economic goals: fostering financial stability; contributing to good macroeconomic performance; and securing efficient patterns of financial intermediation. Politically, agencies responsible for these policies (regulators) simultaneously nurture their own bureaucratic self-interest, sometimes by attempting to intervene "advantageously" in the sectoral allocation of credit. Conflicts arise not only among these four goals at a given time, but also across time with respect to the intended and unintended effects of particular policies.

Policymaking environments may be conceived as sets of evolving economic and political constraints within which an agency's leadership seeks to maximize a stationary objective function. Changes in operative constraints either may be *exogenous* to the agency, or may be the *intended* or *unintended result* of the policies it follows. The technology of regulation embodies concepts and powers whose status is rooted in applicable bodies of law, particularly in constitutionally guaranteed freedoms, in an agency's enabling legislation, and in judicial reviews of its past actions. For a regulatory agency, the analogue of market constraints lies in limits

imposed on the effective exercise of its statutory tools by market adaptation to their exercise. This adaptation attempts to minimize the net regulatory burdens that agency activities place on regulated institutions (regulatees) and their customers. Regulatory avoidance grows out of efforts by regulatees, less-regulated competitors, and the customers they serve to reoptimize their microeconomic behavior conditional on adjustments in operative regulations. Finally, political pressure on regulatory agencies plays a role similar to regulatory pressures faced by regulatees. Political pressure encompasses a range of carrot-and-stick strategies, centering on possibilities for realigning an agency's budget and its statutory domain and technology of regulation. Pressure transmits societal demands for re-regulation by altering agency perceptions of the costs and benefits of technologically and economically feasible settings of its existing statutory instruments.

In stationary equilibrium, regulation would be optimally adjusted to avoidance activity and avoidance activity would be optimally adjusted to regulation. This general equilibrium perspective informs my title. In principle, structural changes in financial markets alter the constraints under which regulators of financial institutions (and the political constituencies that support their activities) operate. Recent changes in the constraints facing regulatory agencies and their supporting constituencies imply an irresistible political demand for these parties to reoptimize.

## I. The Regulatory Dialectic

Far from emphasizing conditions or stationary equilibrium, my writings on this subject focus on the *dynamics* of regulator-regulatee action and response in a nonstationary context. The guiding idea is an interpretive vision of counterpoised economic and political power that I call the "regulatory dialectic" (see my 1981 article). When we think

\*Everett D. Reese Professor of Banking and Monetary Economics, The Ohio State University, and research associate, National Bureau of Economic Research. Opinions expressed are my own and should not be construed to represent those of the NBER.

of a regulatory action as initiating the process, we may label the three stages in the adaptive sequence as regulation, avoidance, and re-regulation. When structural changes in regulated markets kick off the game, the sequence becomes one of innovation, re-regulation, and avoidance. In both sequences, two critical elements exist: 1) a conflict between creative and hard-to-forecast economic efforts to undo the effects of regulatory activity and political efforts to assert or reassert regulatory control, and 2) the hypothesis that the second or third stage of any given sequence may also be interpreted as the first stage of a new sequence.

In the regulatory dialectic, political processes of regulation and economic forces of avoidance adapt continually to each other like riders on a seesaw. This alternating adaptation is not continuous. Rather it develops as a series of lagged responses. Moreover, because of essential differences in the capacity for creative adaptation (i.e., in the *adaptive efficiency*) of regulators, regulatees, and unregulated competitors, avoidance lags tend to be shorter than regulatory lags.

## II. Contemporary Financial Innovation and the Need for Re-Regulation

Financial Innovation is impelled by regulated and unregulated institutions' adaptation to observed changes in their technological, market, and regulatory constraints and by regulatory adaptation to ensuing changes in regulators' own opportunity sets. Ongoing shifts in technological and market constraints set the problems of re-regulation that confront policymakers today.

### A. Changes in Technological Restraints

Technological advances are increasing the convenience of customer access to financial products and lowering the costs to institutions of providing that access. These changes may be described as the *robotization* and figurative *electronic wiring* of the system for delivering financial services to households and firms. Contemporary high technology is not just replacing letters by electronic messages and thereby eliminating flows of paper

evidences. It is taking people increasingly out of wholesale and (now) retail financial processes. Their place is being taken by computer terminals and software activated by customer plastic cards or telephone calls, and by on-line information files that are interchanged between institutions by cable or satellite communications links.

Robotizing and wiring financial-services delivery systems has two immediate effects: 1) to create *scope economies* that—by spreading joint costs over a wider range of products—allow packages of complementary financial services to be produced more cheaply in combination than singly; and 2) to drive financial transactions costs steadily closer to zero. These changes in cost structure are narrowing deposit-institution lending margins, extending the product lines and organizational structures of financial-services firms of all types, and broadening the geographic area over which existing firms may profitably operate. In financial-services pricing, distance between would-be transactors is increasingly less important. In combination with the holding-company form of organization, electronic technology is shifting the focus of financial institutions from local to national and international clienteles and from stand-alone deposit and loan products to service packages that share facilities and customers with other financial-services firms.

### B. Changes in Market Constraints: *Volatility of Inflation and Interest Rates*

Interest rate volatility has fostered the development of futures and options instruments and markets that facilitate the shifting of portfolio risk. It has also increased (through bracket drift) demands for tax-sheltered investment opportunities and transformed asset-liability *maturity mismatching* by deposit institutions from a reliable source of average profit into a frequent source of substantial losses. Because deposit institutions could pass much of their maturity risk through to federal deposit-insurance agencies, increases in interest and interest risk positioned deposit institutions politically in ways that progressively undercut their long-run share of the savings market.

Interaction between fluctuating market interest rates and deposit-rate ceilings greatly influenced the speed and geographic spread of the robotization and electronic wiring of the system for delivering financial services to households and firms. Given a system of explicit ceilings on deposit interest rates, the rate of diffusion of electronic-funds innovations has varied with movements in the level and term structure of expected future interest rates. Whether or not explicit rates on commercial-bank checking accounts are subject to ceilings, deposit institutions typically offer some forms of implicit interest. This occurs because implicit interest can be used to discriminate quietly among customers and because for households implicit interest is a tax-sheltered form of income. When the excess of market rates over explicit returns on deposits falls below the cost of subsidizing account services, institutions cannot afford to expand electronic-funds services on an unpriced basis. Whenever it was not profitable to increase the rate of implicit interest on deposit balances, the robotization and electronic wiring of the financial system slowed noticeably.

### *C. Changes in Market Constraints: Entry of Less-Regulated Contestants*

Changes in technology and in interest rate volatility change the profitability and risks of financial intermediation. In an exogenously changing economy, government regulation acts in part as an inertial force. Regulation reduces the adaptive capacity of regulated institutions by narrowing the set of economically feasible responses to changing opportunities. Important differences in adaptive efficiency exist between regulated private firms, their less-regulated competitors, and regulatory agencies. Whether changes in the financial environment are favorable or unfavorable, managers of less-regulated competitors of deposit institutions can on average adapt to these changes more quickly and more efficiently than managers of regulated commercial banks and thrift institutions. In turn, deposit-institution managers show on average greater adaptive efficiency in finding ways to circumvent traditional restrictions

on their pricing, location, and business activities than regulators show either in designing regulation or in closing loopholes.

The rapid emergence of nontraditional competitors in financial-services markets becomes more intelligible when we hypothesize three lags in the diffusion of financial innovations. These lags need apply only on average, not for every individual innovation or for every entity involved: 1) the average lag in innovation by less-regulated institutions behind changes in technological and market opportunities; 2) the average lag in innovation by regulated deposit institutions behind their less-regulated competitors; and 3) the average lag of regulatory responses behind innovations of either type.

The first lag reflects delays in project evaluation and gestation. I attribute the lag of regulated financial firms behind less-regulated ones to the self-selected aggressiveness of unconventional entrants into tightly regulated financial markets and to conscious and unconscious resistance to change by managements and employees at many of the established competitors. I trace the third lag to asymmetric information, to the short horizons of political appointees, and to the non-profit orientation and characteristic slowness of bureaucratic decision making.

### **III. Macroeconomic and Microeconomic Issues in Re-Regulation**

In an era of technological, market and regulatory upheaval, deposit-institution managers and regulators must plan for change. During the 1970's, innovation occurred against a backdrop of accelerating inflation. Avoidance of restrictions on product lines, takeover activity, office locations, and deposit pricing was directed against laws that remained framed in terms of technological assumptions appropriate to a bygone era. Technological change occurred, especially rapidly because of its ability to undermine an inherited system of cartel-like segmentation of competition: 1) between deposit institutions in different states; 2) between bank and nonbank deposit intermediaries; 3) between securities firms and deposit institutions; and 4) between financial firms and

nonfinancial businesses. In the 1980's, institutions must worry about disinflation as well as inflation, and must expect regulators to make a concerted effort to erect an architecture of regulation that comes to terms with legally disruptive concepts of communications and information technology.

Macroeconomically, the dilemma focuses on the selection of intermediate targets for monetary policy. Regulation-induced proliferation of deposit substitutes and of schemes for paying implicit interest create a need to define statistical counterparts to concepts of money and its opportunity cost in an evolutionary manner. But difficulties in measurement only partly explain recent switches in monetary policy targets. This is because policy targets externalize political as well as economic goals and constraints. When the Fed adopts targets (as it did in October 1979) whose full implications are unclear to unsophisticated (i.e., adaptively inefficient) constituencies, it temporarily reduces these sectors' ability to use political pressure to protect their economic interests. However, as the distributional consequences of unfamiliar targets unfold, the information imbalance redresses itself. As time passes, the Fed becomes increasingly unable to disclaim responsibility for unpleasant movements in politically sensitive variables such as interest rates. Increasingly, defensive political pressure from these sectors circumscribes Fed policy choices again.

Microeconomically, the dilemma concerns how to rebalance the rights and interests of traditional and nontraditional financial-services players. Regulators can make the framework of competition fairer either by shackling less-regulated players or by unshackling deposit institutions. If considerations of efficient production were all that mattered, the unshackling strategy would clearly dominate. However, political concerns and responsibilities for improving the performance of the national economy and bolstering the stability of the financial system complicate the problem.

Inherited paradigms for regulating deposit-institution markets are poorly adapted to the ways in which competitive conditions are changing. These paradigms treat market

structure as an exogenous determinant of competitor performance and define structure in terms of active competitors in a narrowly conceived line of financial services in a *localized* geographic market. Great emphasis is placed on preventing horizontal combinations that would concentrate a large share of the market in the hands of a few firms. Although banking regulators and the Justice Department have begun to take account of nonbank deposit institutions in the local area and potential entry by other in-state banks, the Supreme Court has so far rejected this view.

Specialized regulators have been quicker to see the need for adapting the structure-performance paradigm and have done so implicitly by allowing holding-company acquisition to develop as a substitute for branches and mergers. However, industrial organization in research published by Fed staff members has lagged, especially in recognizing the contribution of money market funds, brokerage cash-management accounts, and affiliates of out-of-state holding companies to the vigor of *credit-market* (as opposed to *deposit-market*) competition in traditional market areas.

Entry costs are declining to zero and *lending margins* are narrowing as real information and transactions costs fall. Deposit-institution markets are becoming more contestable in William Baumol's (1982) sense of the term.

Contestability theory portrays market structure as adapting, through entry and exit, to minimize *average* production and delivery costs. This theory portrays operating costs, product lines, market structure, and pricing performance as *simultaneously determined* endogenous variables. In particular, it provides a framework that explains how opportunities to reduce the costs of producing packages of financial services would lead to a homogenization of functions across traditionally distinct types of financial institutions.

Omnibus deposit-institution legislation passed in 1980 and 1982 has begun the process of consolidating regulatory functions across agencies and re-regulating deposit interest rates, reserve requirements, institutional activity restrictions, and the pricing of

correspondent services by the Federal Reserve. This legislation implicitly recognizes that it is a losers game to levy implicit excise taxes on producers whose products have (thanks to modern technology) an unlimited number of potential substitutes.

But at the same time that reliance on implicit pricing for deposit and reserve balances is lessening, implicit pricing is playing a larger role than ever in the provision of federal deposit insurance. With deposit insurance explicitly underpriced, the risk exposure of insured institutions can only be controlled by regulatory penalties (i.e., by developing a system of risk-sensitive implicit insurance premiums). However, because avoidance lags are shorter than regulatory lags, some forms of *unregulated risk* always exist, especially in a world of rapid technological change. Interest-volatility and sovereign risk may be recognized as prime examples of risks that went largely unregulated in the recent past. Because unregulated risks are mispriced, insured institutions have an incentive to pursue them energetically. Failure to rationalize the pricing of federal deposit insurance has interacted with technological change and interest volatility to increase the fragility of the entire financial system. The growing threat of worldwide financial crisis challenges regulators and politicians to find a way to price deposit insurance rationally.

#### IV. Summary

Into their conceptions of how policy instruments work, authorities need to incorporate the dialectical economic and political responses of regulatees and their less-regulated competitors. Until the concept of regulation-induced innovation begins to play a major role in policymakers' analysis of the effects of alternative forms of innovation-induced re-regulation, the possibility of financial instability remains a serious threat.

#### REFERENCES

- Baumol, William J., "Contestable Markets: An Uprising in the Theory of Industry Structure," *American Economic Review*, March 1982, 72, 1-15.
- Kane, Edward J., "Accelerating Inflation, Technological Innovation, and the Decreasing Effectiveness of Banking Regulation," *Journal of Finance*, May 1981, 36, 355-67.
- , "Selecting Monetary Targets in a Changing Financial Environment," in *Monetary Policy Issues of the 1980s*, Kansas City: Federal Reserve Bank of Kansas City, 1982.

# Financial Innovation in Canada: Causes and Consequences

By C. FREEDMAN\*

From November 1975 until November 1982, the Bank of Canada set explicit targets for the narrow monetary aggregate, *M1*, which is composed of currency and net demand deposits (i.e., demand deposits net of float). The latter category, in turn, comprises personal chequing accounts and current accounts with the former used solely by individuals and the latter used primarily by businesses. Although there had been financial innovations in the late 1960's, largely as a result of the 1967 revision of the Bank Act, these had no significant direct effects on policy since the principal intermediate target of policy in those years was the level of nominal interest rates. However, by removing ceilings on administered rates in Canada, the 1967 Bank Act removed one significant potential source of financial instability and innovation, namely the development of new instruments outside the banking system that would be able to attract funds from the banking system because of the constraints and restrictions placed upon the latter. Indeed, it is worth stressing that Canadian experience with respect to financial innovation over the last seven years has been entirely the result of market factors—competition, technological innovation, and high interest rates—as opposed to the kind of innovation related to deregulation that has occurred in the United States and elsewhere.

Since the beginning of monetary targeting in 1975, there have been four specific episodes of financial innovation that have had significant impact on the demand for *M1*, two relating to the household sector and two relating to the corporate sector. I will first set out in some detail the changes that have occurred, and then examine the types of

responses that have been available to the monetary authorities.

## I. Household Deposits

### A. *The Introduction of Daily Interest Savings Accounts in 1979*

Until 1979, the standard savings account in Canadian chartered banks paid interest on the basis of the minimum balance held in the account over the calendar month. As a result, small savers were, by and large, unable to earn income on funds that were available for periods of less than one month, such as salary payments that were received during the month, and therefore tended to keep these funds in their transactions accounts. Following the lead of the near-banks and smaller chartered banks, the large banks began to offer daily interest savings accounts to their customers towards the end of 1979. Interest on these accounts is computed on the daily closing balance, thus offering the small saver the opportunity to earn near-market rates of interest on liquid assets held even for short periods of time. Not surprisingly, these accounts proved very popular and grew very rapidly. Although by far the greater part of the funds deposited in the daily interest savings accounts were shifted from other forms of savings accounts, a small part reflected a reduced demand for personal chequing accounts and hence a decline in the demand for *M1*.

### B. *The Spread of Daily Interest Chequing Accounts Since 1981*

With very minor exceptions, the daily interest savings account could not be used as a transaction account. However, a number of near-banks and smaller banks began to offer daily interest chequing accounts which combined features of both the daily interest savings account and the personal chequing

\*Chief, Department of Monetary and Financial Analysis, Bank of Canada. The views expressed in this paper are my own and no responsibility for them should be attributed to the Bank of Canada.

account. Competitive pressures led to a gradual spread of this "all-in-one" account and most of the major banks introduced it as an option for their customers during 1981 and the first half of 1982. Typically, this account offers a rate just under the rate on daily interest savings accounts on closing daily balances above some minimum balance (\$1,000 and \$2,000 are fairly common), and a much lower rate (3 percent) on daily closing balances below the minimum. Some banks have even more tiers such that different rates are paid on daily balances below \$500, between \$500 and \$1,000, and above \$1,000. A number of financial institutions foresee this type of account as the principal household account of the future, perhaps with sufficient options offered to customers to make this the only account needed for both savings and chequing purposes. Since the daily interest chequing account is technically a "notice" deposit and not a demand deposit, it is not included in *M1* as currently defined.

The most significant factors in the introduction and spread of both new types of household accounts were the diffusion of the technology that enabled financial institutions to offer such accounts, the competitive nature of the financial system (including both banks and near-banks), and the prevailing high level of interest rates. The smaller institutions tended to lead in the development of the new "products" and the larger institutions felt constrained to match them as soon as the computerization of their branches was sufficiently widespread to make it feasible to offer the new type of account. The high interest rate environment of the mid- to late-1970's played an important role in these developments insofar as households became much more sensitive to the interest foregone on their noninterest bearing chequing accounts and, therefore, the potential gains to an institution of introducing the new types of accounts became much greater than at periods of low interest rates. This set of circumstances induced financial institutions to innovate in order to improve their competitive positions. Once the new instrument became available at some institutions, com-

petitive pressures ensured its spread across virtually the entire financial system.

## II. Corporate Accounts

There were two main periods of financial innovation on the corporate side—the mid-1970's and the last couple of years. They differ not so much in the nature of the innovation as in the type of company affected by the changes. In the earlier period, cash management packages were offered to very large companies and to large governmental organizations, while in the more recent period similar schemes spread to intermediate-sized companies. In addition to the direct influence of these packages on the demand for transaction balances, a number of technical changes in the way these accounts are booked and in the way service charges are paid have also affected the demand for *M1*.

Before describing these developments in detail, it is worth noting that the primary factors leading to the spread of cash management packages have been the new technology permitted by widespread computerization, the competitive atmosphere engendered initially by the availability of short-term interest paying instruments in the market, and later by the aggressive attempts to increase market share of some of the medium-sized banks, and the high interest rate environment which increased the cost of holding non-interest-bearing balances. The earlier development of cash management packages in the United States may also have played a role in sensitizing corporate management to the potentialities of making use of idle balances.

There are a number of aspects of corporate cash management packages that bear on the demand for *M1*. First, consolidation into a single account of funds flowing into several, possibly geographically dispersed, accounts has enabled corporations to reduce the level of working balances. Second, the use of techniques such as regional lock boxes and pre-authorized account withdrawals for speeding up inflows and the introduction of techniques such as payroll service plans and con-

signment cheque plans to impose stricter control over disbursements has permitted a further reduction of working balances. Third, and most important for our purposes, the banks have arranged for surplus funds to be used profitably overnight, accepting standing instructions from corporations as to how to employ the funds.

There are a number of different ways in which the overnight investment takes place. In some cases, the banks pay explicit interest on current accounts at a rate that is below, but moves with, the prime rate. In others, the banks impute to these accounts implicit interest which is used to offset both service charges on the operation of the account and float charges relating to the timing of deposits and withdrawals. Some banks tend to use notice deposits, particularly nonpersonal chequeable deposits and nonpersonal non-chequeable deposits (neither of which is included in  $M1$ ) as the repository of these overnight funds. Balances that are likely to be available for more than one day can also be put into short-term nonpersonal fixed-term deposits at the banks and even one-day CDs are available although not in very widespread use for working balances. It is also possible to negotiate automatic pay-downs of outstanding demand loans with the closing balance in the current accounts. Finally, although the banks will automatically implement some or all of the above transfers for their customers, the latter can also use their cash balances to invest in overnight money market instruments, although such a decision must be made earlier in the day and cannot be based on the closing balance.

As I have already emphasized, unlike the situation in the United States, deregulation played absolutely no role in the developments in either the household or the corporate sector. No prohibition has ever existed to prevent banks and near-banks from paying interest on personal chequing accounts or current accounts. The decision of financial institutions to pay near-market rates of interest on some transactions balances resulted from the combination of higher nominal rates of interest and the competitive nature of the financial system. The split reserve require-

ment with a higher ratio on demand deposits than on notice deposits provided the incentive for banks to try to book the interest-paying transactions balances as notice deposits, if necessary creating new instruments in so doing, rather than to pay interest on existing demand deposits. It may also have played a role in the decreasing use of compensating current account balances in favour of explicit service charges.

### III. Policy Implications

The demand for money equation can be written as  $M = f(Y, r, \dots)$ . The crucial element for the policy linkage is a reasonably tight relationship between money, total spending, and the nominal interest rate. Thus, for example, a rapid growth of nominal income will be reflected in a rapid growth of money causing the monetary aggregate to exceed its target. The appropriate policy response is to increase interest rates, thereby causing the monetary aggregate to return to its target, albeit with a lag, and, with a somewhat longer lag, slowing down the growth of nominal income. In Canada, the authorities have used the demand for money equation to determine the appropriate interest rate levels to bring the monetary aggregate back to its target over a period of time sufficiently long to avoid excessively volatile movements of the interest rate. (See my 1981 paper for a detailed discussion of the feedback rule approach to monetary targeting.)

The three dots in the equation represent all other factors that influence the demand for money including random errors. Any change in money demand resulting from a movement in such a factor confronts the authorities with a choice between two alternative responses: (i) allow the quantity of money to change without adjusting interest rates; or (ii) adjust interest rates so as to bring the recorded monetary aggregate back to its target. The former response requires the authorities to make periodic and convincing explanations of why the monetary aggregate is outside its target range; the latter involves an inappropriate response to what is clearly an "autonomous" shift in money de-

mand (i.e., one not induced by income or interest rate movements). It is because of the difficulty posed for policymaking by the effect on money demand of factors other than income and interest rate that the Canadian authorities have sought a monetary aggregate in which there are no nonrandom factors other than income and interest rates entering into the function except for shift variables that can be identified and explained. It should also be clear from this discussion why the response to financial innovation has been to search for new stable money demand equations.

There are two major steps required in dealing with a financial innovation. First, it is necessary to determine that a shift in money demand is taking place. Second, it is essential to measure the size of the shift and to adjust either the target growth range or the definition of the monetary aggregate being targeted to take account of the shift. In recent years at least three different ways of implementing the second step have been used in North America: (i) retaining the existing definition and establishing a new target range that takes into account the size of the shift (for example, downward rebasing in Canada announced in February 1981); (ii) redefining the monetary aggregate in such a way as to internalize the shifts that are occurring (for example, U.S. redefinition of *M1* to include NOW and ATS accounts); and (iii) creating a new artificial monetary aggregate that adds back the fraction of any new instrument that represents the shift out of the previous monetary aggregate (for example, shift-adjusted *M1B* in the United States). I will now turn to a discussion of the Canadian experience of the last six years to throw further light on these issues.

Identifying shifts in the demand for money is substantially easier in the case of deregulation or the introduction of new instruments than in the case of innovation unaccompanied by the creation of a new instrument. Thus in the case of innovations on the household side, which involved the introduction of daily interest savings accounts and daily interest chequing accounts, the authorities were able to identify and track

the new developments from the very beginning. The identification of the initial shift in corporate accounts was difficult since it took place gradually over a rather long period of time. Furthermore, although there was much greater sensitivity to the possibility of a spread of cash management packages to intermediate-sized corporations after the experience of the mid-1970's, there was still a time lag between the beginning of this new development and the identification of its occurrence, in large part because of the fact that the driving forces behind the diffusion of the cash management packages were the regional offices of the banks and not their head offices. The normal flow of institutional information thus did not immediately signal the advent of a new innovation.

The second and more difficult problem is measuring the size of the shift and adjusting the target. One technique used to measure the shift has been to compute the actual or, in some cases, the average errors from a postsample simulation of the existing *M1* equation. A second technique that produces very similar results is to introduce a dummy or transitional variable into the *M1* equation and to use the coefficient on that variable as an estimate of the shift. A third approach that has been used applies the same techniques not to the aggregate *M1* equation, but to the relevant component equation, the personal chequing account equation in the case of a shift in the household demand for money and the current account equation in the case of a shift in the corporate demand for money.

There are a number of problems inherent in these procedures. First, there is a minimum number of data points, that is, a minimum period of time, needed to be able to make judgements about the magnitude of the shift even when institutional information clearly indicates that a financial innovation has taken place. There may thus be a non-trivial lag between identification of the shift and its measurement, and during such periods monetary aggregates are not a very reliable guide for policy. Second, the less precise the preshift equation, that is, the larger the intrasample errors of the equation, the less

firm one can be about the magnitude of the shift on the basis of postsample forecasting errors of the equation. Third, if more than one change is going on at the same time as was the case in the most recent period of innovation, it may be even more difficult to pin down the size of the shifts.

Another problem in the most recent episode was the sensitivity of the estimates of the shift to changes in specification of the equation. With respect to functional form, for example, until recently there has been very little difference between the performance of semilogarithmic and logarithmic equations, and hence no basis on which to choose between them. The crucial experiment which would have allowed one to discriminate between the two equations was the high interest rate environment of the last three years. But this is precisely the period in which the shifts were also taking place. If one uses a semilogarithmic equation, much of the slow growth in  $M1$  of the recent period can be attributed to the level of interest rates with less room left for a shift. On the other hand, with a double-logarithmic specification one explains much less of the slow growth in  $M1$  by the level of interest rates and much more via the shift. Unfortunately, the data are simply not rich enough to yield a definitive conclusion to date.

As mentioned earlier, there are three ways of adjusting the monetary aggregate target—shifting the base and/or range, changing the definition to internalize the shift, and adjusting the definition to take account of the fraction of the new instrument coming from the components of the old aggregate. In the case of both the first corporate sector shift and the introduction of daily interest savings accounts, the reaction of the Bank of Canada was to reduce the base from which the growth of the monetary aggregate was computed, but not the target growth rate itself. This is the appropriate response to a once-and-for-all shift in the constant term of the demand for money equation. Considerable effort has gone into calculating the downward shift that has taken place as a result of the introduction of the daily interest chequing account and the spread of cash

management packages to intermediate-sized corporations, with a view to carrying out the same type of rebasing exercise. The principal reason that this attempt has not resulted in a new target for  $M1$  is that the shift out of  $M1$  was ongoing and not a completed once-and-for-all shift. Therefore one had to estimate both the size of the shift at a given date in the past (in order to choose a new base) and the effect of the shift on  $M1$  over the future (in order to choose a new target growth rate). Difficult as the first part of the exercise proved to be, it was relatively straightforward compared to the problems of predicting future shifts which depended on such factors as the marketing expenditures of the banks, the spread of computer technology, and the prevailing level of interest rates.  $M1$  is, therefore, not considered to be a useful guide to policy at the present time and the Bank of Canada has withdrawn the target range for  $M1$ . (See Gerald Bouey.)

More recently, attention has been focused on the possibility of redefining the aggregate in order to internalize the shift, thus obviating the problem of having to estimate the magnitude of an ongoing shift. Among the primary candidates for inclusion in a new aggregate are nonpersonal chequable and nonchequable deposits and daily interest chequing accounts. An important consideration in interpreting the movements of any aggregate that contains daily interest chequing accounts is that these accounts have both transactions and savings characteristics, and therefore a significant part of their growth may be at the expense of nontransaction deposits. Furthermore, to explain adequately the movements of a broader aggregate, one may have to take into account movements in wealth and the own-rate of interest, in addition to income and competing interest rates. Unless the own-rate is perfectly correlated with market rates of interest, movements in the margin between the own-rate and competing rates may significantly influence the growth of the aggregate in the short run, and the authorities will have to take these developments into account in making policy. More generally, as discussed above, the introduction of variables other than income and in-

terest rates to the demand for money equation may require the authorities to explain why they are not responding to movements in the monetary aggregate outside the target range when it is the movements of these "extraneous" variables that are the primary causes of the aggregate being outside its range.

The third method of adjusting the target, replication of the U.S. approach in creating shift-adjusted *M1B*, is not possible in Canada because of the lack of information needed to carry out such a complex assignment. There are no data currently available on the proportions of daily interest savings accounts and daily interest chequing accounts that have come out of PCAs.

What is the current situation? The work on revised aggregates has not yet been completed. However, it is already apparent that, because of the greater heterogeneity of the components of these aggregates and the re-

lated reduction in the precision of the estimated demand equations for the aggregates, the way in which they could be used will be somewhat different from the way *M1* was used in the past. It may be that any future role of the monetary aggregates will be more in terms of a medium-term check against cumulative error rather than as a guide to short-term policymaking and any use of target ranges would need to reflect this fact.

## REFERENCES

- Bouey, Gerald K., "Recovery from Inflation," Notes for Remarks to the Canadian Club, Toronto, Ontario, November 29, 1982, *Bank of Canada Review*, December 1982.
- Freedman, Charles, "Monetary Aggregates as Targets: Some Theoretical Aspects," Technical Report 27, Bank of Canada, and Working Paper No. 775, National Bureau of Economic Research, 1981.

# THE ROLE OF ALIEN ENTREPRENEURS IN ECONOMIC DEVELOPMENT

## An Entrepreneurial Problem

By PETER KILBY\*

Alfred Marshall's fourth factor of production has proved very hard to nail down. In particular, the economic importance of the entrepreneurship element in "organization," its empirical measurement and its socioeconomic determinants all remain, nearly a century later, in the category of unfinished business. In this paper an attempt is made to outline the current state of this unfinished business with respect to underdeveloped countries—how the problem has been conceptualized, the type of evidence collected, and the conclusion of some that it is indeed not a problem. I propose a slight redefinition of the entrepreneurial task, the type of evidence appropriate to its measurement and, of course, a different conclusion.

In economic theory, entrepreneurship finds a place only in the lower realms, where imperfect knowledge and market failure are granted an untidy presence. Theorists of the lower realm such as Harbison, Hagen, and Harris would agree that entrepreneurs supply services that are unavailable in the market place, but what exactly these services are would be disputed. Among those who study the world of men, economic historians of the now developed countries are in accord, even if few of them are included to study it, on the vital role of the entrepreneur in the development of these economies. But as to the notion of a potentially inelastic supply, a "vital few," these scholars divide, with a growing segment adhering to the agnostic view (see, for example, Douglass North, p. 8).

In underdeveloped economies the entrepreneur, as a potentially critical bottleneck, has been given a more prominent place. On the one hand, the residual tasks that fall to the business leader are far greater when well-integrated markets, particularly financial and factor markets, are lacking and when decision making is impeded by the consequences of incipient political instability. On the other hand, the number of individuals capable of fulfilling these tasks appears to be sharply circumscribed by an inheritance of deprivation for sizable segments of the population, by narrow commercial traditions, and by brief exposure to "technological culture."

### I. An Entrepreneurial Problem Denied

Despite this strong a priori expectation, empirical support for inelastic entrepreneurial supplies has not been forthcoming. Most studies have been cross sectional, based on entrepreneur interviews of several hours duration. Performance (size of assets, sales, employment, profits) has been related to education, occupational background, social status, political influence, religious affiliation, ethnic origin, and various economic variables. No attempt has been made to track psychological characteristics. When the effect of other variables is accounted for, ethnic and religious factors have had little influence on performance, suggesting that the potential contribution of all segments of the population is available to the entrepreneurial pool. Although education, rather surprisingly, has little discernible effect, business performance is responsive to economic variables, as well as to access variables (social status, political influence). Investigators have tended to interpret these findings as indicating the ab-

\*Professor of economics, Wesleyan University. I have benefited from the advice of Stanley Lebergott and Michael Lovell.

TABLE 1—OUTPUT GROWTH RATES IN LESS DEVELOPED COUNTRIES  
(Shown in percent)

		GDP		Manufacturing	
LDC Market Economies	1950–59	4.6		6.9	
	1960–73	5.7		7.5	
Low-Income Economies	1960–70	4.4		6.5	
( $\bar{Y} < \$430$ )	1970–80	3.5		3.6	
Middle Income Economies	1960–70	5.9		6.8	
(\$440 < $\bar{Y}$ < \$4,500)	1970–80	5.6		6.4	
Per Capita Manufacturing Output Growth Rates, 1970–80 (selected non-OPEC countries)					
Ethiopia	0.4	Jamaica	–3.7	India	2.9
Malawi	3.8	El Salvador	1.2	Turkey	3.9
Mozambique	–9.8	Columbia	4.3	Burma	2.0
Sudan	–1.7	Bolivia	3.5	Taiwan	10.3
Ghana	–5.9	Peru	0.6	Thailand	8.1
Kenya	8.0	Chile	–2.2	Korea	14.9
Ivory Coast	2.2	Brazil	8.1	Singapore	8.1

Source: First two rows from UN *Yearbooks of National Income Statistics* as recorded by Leff; all other data from World Bank, *World Development Report 1982*.

sence of a significant entrepreneurial problem.

However critical limitations of research design should be noted. In the absence of reliable records, performance variables are not infrequently suspect. Further, large assets or sales can be the result of an undisclosed “government factor” (government loads, sales to government agencies), with the former not surviving the withdrawal of the latter. More fundamentally, performance is a censored variable with no observations for those entrepreneurs who fail, and none for the upper reaches of the manufacturing sector where alien minorities, multinational corporations, state, and joint enterprises predominate.

In a provocative essay, Nathaniel Leff has recently argued on quite different grounds that the entrepreneurial problem for backward economies has been solved. Citing the statistics shown in the first two rows of Table 1, he contends that high growth rates are *prima facie* evidence that entrepreneurship has not imposed a constraint on economic development. The actions of political leaders and policymakers reflected in the “restrictions, expulsions and massacres” of alien minorities, Leff contends, indicates their

confidence in an elastic supply of indigenous entrepreneurship. Signalling a scholarly reappraisal, this subject has received little attention in journal articles and textbooks since 1970. Leff argues that the entrepreneurial constraint has been released and this as a result of (i) government pricing policies that have raised returns and reduced risks, (ii) path-breaking activities of state enterprise, and (iii) the formulation of zaibatsu-like “groups” in the private sector.

Leff's argument is unconvincing. Aggregate statistics on output growth hardly provide useful evidence. If domestic entrepreneurs are not responsible for the bulk of output in a country's manufacturing sector, high rates of growth cannot be taken as an indication of their supply elasticity. If, assuming that domestic entrepreneurs are dominant, one has recourse to undifferentiated national growth rates, certainly the first step ought to be an examination of variance among countries, with an eye to those economies thought to be poorly endowed and those well endowed with entrepreneurial resources. Since the focus is on development, the proper measure is growth in *per capita* manufacturing output. Finally, these figures should be adjusted downward, removing the inflation

of value-added introduced by protective subsidies; statistics reported by I. M. D. Little et al. suggest a halving of reported growth rates.

Average reported growth rates shown in Table 1 have come down in the past decade. The advance in per capita manufacturing output was 5.4 percent for middle-income OPEC economies (entrepreneurial agents here are *MNCs* and joint enterprise), 3.9 percent for other middle-income economies, and for countries with a per capita income below \$430—where a priori considerations would predict greatest entrepreneurial constraint—the figure is indeed a low 1.6 percent. Among individual countries, variation is marked. Of the six countries with growth rates of 8 percent or more, four are economies where Korean or Chinese entrepreneurs predominate. All in all, it is hardly a strong case for the universality of elastic supplies of entrepreneurship.

## II. An Entrepreneurial Problem Revealed

High rates of output growth can be taken as evidence for the absence of an entrepreneurial bottleneck *if* effective rates of protection are under 30 or 40 percent, and *if* the additional output is coming from domestic entrepreneurs. Neither is usually the case. What is seldom noted in evaluating the entrepreneurial issue is that private domestic entrepreneurs, after a quarter century of industrialization, still account for a small fraction of manufacturing output. In Ghana, the share is 6 percent, in Ivory Coast, it is 11 percent, and in Malaysia, it is about 5 percent. Of the two other high-growth countries in Table 1, African entrepreneurs account for only 9 percent of recorded value-added in Kenya; in the case of highly developed Brazil, private businessmen contribute only 27 percent. In these and most *LDCs* the bulk of the entrepreneurial talent is supplied by *MNCs* (for example, 66 percent in Kenya, 51 percent in Brazil), state enterprise and alien minorities. Various forms of joint enterprise aside, in 100 percent state-owned and operated ventures key technical and managerial positions are frequently filled by foreign per-

sonnel. In virtually no country—whether it be Ghana or Korea—have state-managed industrial enterprises proved economically viable.

A second feature of the industrial structure of most backward economies is a very large number of domestic firms, in the informal sector and the bottom rungs of the enumerated manufacturing sector, employing from one to twenty workers. There seems to be a barrier beyond which these entrepreneurs cannot expand. Of the establishments which are larger, a good many are the result of subsidized government lending schemes and nursery industrial estates; failure rates run in the 30 to 50 percent range, with very few of the successes going on to expand beyond their initial employment levels of twenty to fifty workers.

From these facts, what can we infer about the nature of the entrepreneurial task in late-developing countries? Most writers have stressed the perception of market opportunity, the ability to make investment decisions, to innovate, to bear risk. Clearly the evidence from industrial structure and the field investigations summarized earlier demonstrates these attributes are available in abundant supply and do not constitute a bottleneck. It is otherwise with the day-to-day functions of managerial coordination and production control, tasks which economists have been reluctant to classify as entrepreneurial since Schumpeter. In import-substituting, technology-borrowing countries, Schumpeter's innovator vs. mere manager priority must be reversed. It is precisely inefficiencies in the routinized managerial functions that prevent domestic entrepreneurs from continuously expanding their firms and from moving into more complex manufacturing activities.

I first observed these deficiencies as being the fundamental bottleneck to advancement in scale and technological complexity in Nigeria in the early 1960's; my 1962 paper presenting these data, along with an analysis of ILO productivity mission results for the 1950's, provided the empirical foundation for Harvey Leibenstein's 1966 "X-Efficiency" article. Recent field work in Central America, East Africa, and Southeast Asia

convinces me that the problem endures, despite rising levels of education and extensive government assistance.

Stated most simply, when one visits the firm of a native businessman, an entrepreneurial problem usually manifests itself in a slow rate of throughput when production is under way, as compared to firms operated by alien entrepreneurs utilizing the same equipment. Moreover interruptions in production are frequent, quality suffers from variability in product specifications, and there are substantial leakages in terms of raw material wastage, pilferage and clerical embezzlement. These can be related to a low degree of coordination and planning, a disinclination to utilize written records intensively for purposes of control, and the absence of conscientious supervision in the workplace. The consequence of these shortcomings is to penalize net earning, to limit the competitive strength of the firm, and to prevent the entrepreneur from moving into more-demanding productive activities.

The following systematically sets out the managerial variables, and their determinants shown in parentheses, which control the technical efficiency with which the factors of production are combined within the firm.

$$Q = \alpha L^{\lambda_L} K^{\lambda_K} M^{\lambda_M}; \alpha(D, E, T, W, Q),$$

where  $D$  = Designed capacity (Embodied in equipment);  $E$  = Efficiency (Plant layout, work methods, materials handling, supervision);  $T$  = Time: Intrashift utilization (Maintenance, control of stock and spares, supervision) and Shifts per day (Planning, willingness of supervisory staff);  $W$  = Wastage: Materials wastage (Work methods, supervision, inventory control) and Theft (Inventory control, monitoring accounts);  $Q$  = Quality (Work methods, supervision).

Why is there low-intensity performance in these particular activities whereas other entrepreneurial tasks are well executed? I have speculated elsewhere (1971) on the sociological antecedents that seem to explain differential role performance. And, of course, economic policy variables powerfully influence learning over time. Particularly relevant here is that subset of policies which affect the

contribution alien minorities make to this process.

### III. The Role of the Alien Entrepreneur

Compared to the native population, alien minorities are at an advantage in the entrepreneurial sphere for a number of reasons. First, the Levantines in Latin America aside, they have possessed a superior initial endowment of capital, market and technical knowledge, and acquired traditions. Second, external environmental parameters—limited occupational choice, the intensifying effect of the never-distant threat of expulsion, greater freedom to undertake extra-legal arrangements—tend to strengthen entrepreneurial performance. Lastly, enforced cooperation with fellow aliens as the sole path to survival builds up over time networks of trust which provide access to scarce information, to various risk-spreading arrangements, to credit on favorable terms, to influential people, and to a larger pool of individuals to whom portions of managerial responsibility can be safely delegated.

While domestic entrepreneurs cannot replicate these "advantages," they can learn the management practices directly by working for or in collaboration with minority businessmen. Indeed, apprenticeship with aliens has been a far more important seedbed for medium-scale domestic firms than association with the much-studied *MNCs*. Starting from a very small size and a simple labor-intensive technology, these alien minority firms have grown under the same conditions of restricted access to finance and specialized personnel that native businessmen face. They have overcome these constraints by capital-saving, cost-reducing innovations that entail (a) substituting carefully selected domestic raw materials for prior imported materials, and (b) replacing machine-controlled product quality and work pace by a system that combines close supervision and payment-by-results.

There are two types of policy changes needed to release the full potential of the aliens' contribution. The first is simply the removal of the *de facto* discrimination against these minorities with respect to licensure,

enforcement of government regulations and access to development assistance programs. The second set of policy changes are those that encourage rather than inhibit the formation of private joint ventures, currently an under-the-table preserve of politicians. Full disclosure, monitored sharing of managerial responsibilities and provision for a buy-out at appraised market value after ten years are some of the ingredients for a successful "transfer of technology."

#### REFERENCES

- Kilby, Peter, "Organization and Productivity in Backward Economies," *Quarterly Journal of Economics*, May 1962, 76, 303-10.
- \_\_\_\_\_, "Hunting the Heffalump," in his *Entrepreneurship and Economic Development*, New York: Free Press 1971.
- Leff, Nathaniel, "Entrepreneurship and Economic Development: The Problem Revisited," *Journal of Economic Literature*, March 1979, 17, 46-64.
- Leibenstein, Harvey, "Allocative Efficiency vs. X-Efficiency," *American Economic Review*, June 1966, 56, 392-414.
- Little, I. M. D., Scitovsky, T., and Scott, M., *Industry and Trade in Some Developing Countries: A Comparative Study*, London: Oxford University Press, 1970.
- North, Douglass C., "The Economic Growth of the United States 1790-1860, Englewood Cliffs: Prentice-Hall 1961.

# Chinese Entrepreneurs in Southeast Asia

By YUAN-LI WU\*

To the student of alien entrepreneurship and economic development, Chinese businessmen in Southeast Asia offer a rich case study. The history of their experience supplies some interesting answers to questions raised in the literature (see Everett Hagen, Bert Hoselitz, Peter Kilby, John Kunkel, and D. C. McClelland): such as the elasticity of supply of entrepreneurship; the origin and motivation of alien entrepreneurs; their changing functions and spheres of action; their impact on local entrepreneurship and the host country economy; the transmutability of their dynamics, etc.

The earlier activities of Chinese entrepreneurs in the region traced a pattern of initial dominance in retailing and finance, followed by greater participation in manufacturing. Their impact on the host country economies and on the emergence of native entrepreneurs has not, however, been uniform. During the last decade and a half especially, this impact has varied with the degree of their acculturation, the host country's policy of acceptance or exclusion, and the elasticity of supply of native entrepreneurship. To the extent the Chinese have remained as alien entrepreneurs without being fully accepted, their function has been strongly affected by the adverse sociopolitical environment. Yet some of the side effects of their response to the environment are far from being negative. The evidence to date described below impinges on more than one theory and has some far-reaching policy implications.

## I. Historical Background

According to Shozo Fukuda, Chinese investment in Southeast Asian countries in 1930

was heavily concentrated in trade and services, varying between a low of 67 percent in Indonesia and a high of 92 percent in Thailand. Exceptions were Malaya and Singapore; owing to their heavy involvement in rubber at that time, nearly one-half of Chinese investment was in the primary sector. In Indonesia, Chinese sugar investments raised the corresponding ratio for the primary sector to 30 percent. Chinese employment in the region was then less heavily concentrated in the tertiary sector than was Chinese capital; a sizable portion of the Chinese work force was in manufacturing.

After World War II, when the British and Dutch colonial regimes had departed, Chinese enterprises were clearly dominant in domestic and intraregional trade in Malaysia, Singapore, Indonesia, the Philippines, and Thailand. But for a spate of anti-Chinese laws and regulations in all Southeast Asian countries except Singapore, the trend in the subsequent two decades pointed unmistakably toward even greater Chinese economic dominance.

Acculturation and intermarriage, with mutual interaction, had been long at work in reducing the alienness of Chinese in Thailand and the Philippines, thanks largely to religious tolerance in the two countries. In both cases, intermarriage has gone very far and bilingualism is quite common. The absence of any blockage to political and social advancement by Chinese in Thailand is especially remarkable. With virtually complete assimilation, it will become increasingly difficult to identify Chinese economic activities in Thailand as a separate category. In time, the same will probably happen in the Philippines.

In Indonesia and Malaysia, the native political authorities have continued to discriminate against their ethnic Chinese citizens on racial and religious (in Malaysia only) grounds. Intermarriage and assimilation are correspondingly much lower

\*The University of San Francisco. This paper is dedicated to the late Professor Chun-hsi Wu, a friend and co-worker on Southeast Asia. I wish to thank the Hoover Institution for a research-travel grant in summer 1982.

(negligible in Malaysia). However, ethnic Chinese businessmen are now back in Indonesia's national retail network. They dominate commercial banking and play an important role even in the state banks. The services of Chinese agents are sought after by Western and Japanese investors attempting to form joint ventures locally. Chinese businessmen have become virtually indispensable to the political power holders themselves. Even in Malaysia where Chinese entrepreneurs have had a difficult time, some have continued to prosper. Their exclusion from the military, the police force, and higher-ranking positions in the civil service have added to the incentive for economic pursuits.

## II. Characteristics of Early Chinese Economic Activities

First, the Chinese immigrants were a hardy, self-reliant and, above all, risk-taking lot. One group of early arrivals consisted of traders who, while plying their commerce with Arabia and India by sea, went to the "South Sea Islands." Others went from the thirteenth century onward as refugees from numerous natural disasters and dynastic wars in China. Later, many Chinese were recruited as plantation and mine workers by the British in Malaya and the Dutch in the East Indies. While some were "Shanghaied," the majority probably chose to go on their own. There were no real Chinese government-sponsored efforts to send workers or settlers to the region. Emigration from China was prohibited by the imperial government for extended periods, and at no time could emigrants expect governmental assistance. Thus those who went to Southeast Asia did so on their own and at their own risk. These circumstances constituted in effect a strenuous process of selection on the supply side of Chinese emigration.

Second, the Chinese immigrants came from an economy that was highly developed commercially. Trade and agricultural credits, deposit banking, and remittance were familiar practices. Since the native populations were largely subsistence farmers or fishermen, the Chinese became traders among them, acting

as the innovating, deviant entrepreneurs of theory. Later, linguistic ability and familiarity with local conditions enabled the Chinese to serve as intermediaries between native producers and Western buyers for export, and as distributors of imported manufactures nationwide. They provided agricultural and trade credits that would otherwise have been lacking. As craftsmen or even laborers they constituted a labor force that readily responded to industrial discipline and training, thus filling another gap the colonial administrators and expatriate businessmen faced. A strong demand existed for Chinese labor and entrepreneurship in the distributive, mining, and light manufacturing sectors.

Third, the Chinese possessed qualities of traditionally Chinese and Confucian origin, not shared by the native populations, that made the progressive expansion of their business activities and advance up the technological and economic ladder a natural process. These were 1) a high propensity to save and to reinvest business earnings, 2) a universally strong desire to secure a better education for their children who would then be expected to carry on the business and often did (a high *n*-achievement), and 3) a strong sense of loyalty and mutual obligation within the Chinese "extended family." The first quality contributed to internal financing; the second to the gradual upgrading of business and production methods and expansion into multiple industries by members of the same family. Having often begun as manual workers themselves, Chinese entrepreneurs had no aversion to "dirtying their hands" in production. The third quality meant wide use of credit and cooperation in marketing, and credit extension among members of the same family association or speech group. Personal trust supplied the basis of the verbal contract which soon became an integral part of Chinese business style. In the absence of enforceable legal protection, it was a condition for survival. Although biases in favor of family members could lead to nepotism while business cooperation within a particular group could lead to unfair competition vis-à-vis outsiders, the effects of these behavioral traits appeared to have been more positive

than restrictive in Southeast Asia before World War II.

### III. Political Disinterest and Its Economic Consequences

If economic qualities alone had been the decisive determinants of their fate, the characteristics exhibited by Chinese entrepreneurs and workers outlined above would have made them the obvious contenders to replace the Westerners in postwar decolonization. Yet in countries where they were a minority, and today notably in Indonesia and Malaysia, ethnic Chinese have become targets of xenophobic—in some instances racial—discrimination. Even in postindependent Malaysia where the Chinese account for 35 percent of the population and could have competed effectively for political power, the realization that economic ascendancy was not enough to save their skin came too late. Chinese economic dynamism has failed to carry over into the political field.

The ethnic Chinese were disinterested in local politics before, and for many years after, World War II because many regarded the countries of their lifetime economic activity as places of “temporary” sojourn. The sentimental and moral bond of the Chinese extended family and ancestral system gave them a dream of retirement in comfort back in China. They also had no desire to become second-class citizens in Western colonies. On the other hand, not especially interested in overthrowing the colonial governments either, they were unprepared to join the native nationalists in taking over the government when the former colonial regimes ended.

In dealing with the colonial governments, the Chinese attitude was to be economically useful, and to accommodate themselves to official demands so that they could continue to pursue their own economic activities at minimum cost. By taking the political environment as given, they in turn presented no threat to the colonial administrations. Since the latter authorities were themselves not particularly interested in promoting entrepreneurship and administrative ability among

the native populations they governed, the Chinese entrepreneurs felt no compulsion to show on their part any special social responsibility. To do so might even have earned them less than gratitude from the colonial masters.

Economically, the Chinese entrepreneurs busied themselves with their own affairs. Those who were employees of others strove to become self-employed and, in their turn, employers. Their dream of returning “home” in retirement, often unfulfilled, was probably enough to make them save more than they might otherwise have done. Regular remittances for family support and investing in China became widely known practices. These circumstances earned for the Chinese of Southeast Asia the reputation of being “economic men” and, in the eyes of nationalists and ideologues, alien exploiters, objects of envy and hostility.

By the mid-1970's, however, Chinese entrepreneurs were rapidly revising their self-perception as sojourners. They had become disillusioned with the record of communist rule in mainland China and its treatment of Chinese returned from abroad. Unfortunately, just when most Chinese in Southeast Asia were spiritually no longer China-bound and would rather stay where they were, only the Thai and Philippine governments were sufficiently confident to grant them equal treatment when the citizenship barrier came down generally. In Indonesia, on the other hand, where the Chinese were numerically a tiny minority and have proved virtually indispensable, attempts at assimilation have continued along with discriminatory treatment. In Malaysia, where assimilation has been least and the Chinese are too numerous, the government's present policy seems to be one of “separate but unequal” treatment.

Beginning from the 1950's, both Indonesia and Malaysia have pursued economic policies that aim at redistributing income and wealth in favor of their respective *native* populations at the expense of the Chinese minority, and fostering the growth of native entrepreneurship (see my study with Chun-hsi Wu). Redistribution through “indigenization” has been an avowed objective of Presi-

ident Suharto's New Economic Policy since 1967, and in Malaysia's successive five-year plans. The measures the two countries have adopted fall essentially into four categories: 1) restrictions that prohibit the Chinese from operating certain types of businesses (for example, retail trade) or limit the number of Chinese undertakings; 2) measures that deliberately reduce the profitability of Chinese businesses by (a) employment quotas for non-Chinese native workers who often are less productive, (b) requiring the employment of non-Chinese natives in executive positions and the transfer of stipulated proportions of equity to indigenous owners, and (c) requiring subcontracting with native enterprises; 3) direct subsidization of native enterprises through government contracts and government-financed bank credit; 4) purchase of private corporate shares with tax money by public trusts for subsequent reissue to native buyers in the form of mutual fund shares.

The results of the foregoing programs have been disappointing. The outright prohibition of certain Chinese economic activities has often created serious shortages because the expected native replacement has not been forthcoming. Chinese enterprises have therefore been allowed to return. Attempts to foster native entrepreneurship through preferential contracting and cheap loans have been stymied by the dearth of dependable suppliers, worthwhile projects, and good credit risks. The incidence of bad debts has been high among indigenous borrowers. In a Malaysian effort to increase native ownership of corporate stocks through publicly sponsored mutual funds, native stockholders interested in quick gratification through increased consumption often sell their holdings to Chinese buyers at a small profit. These acknowledged stumbling blocks to indigenization stem essentially from the inability of the indigenous elements to respond positively to the special opportunities offered. Perhaps wrong methods of encouraging native entrepreneurship have been chosen. Perhaps the supply of native entrepreneurship is inelastic and not amenable to manipulation in the short run. This also suggests that the

activities of Chinese entrepreneurs during the colonial period were not a major factor of the lagging appearance of native entrepreneurship.

#### IV. The Chinese Response

There remains the group of measures that, if fully implemented, would seriously alter the ownership structure, decision-making mechanism, and mode of operation of Chinese enterprises. How have Chinese entrepreneurs responded to them?

##### A. *Diversification by Industry*

Obviously, where outright or partial restrictions have been imposed against them, it would behoove the Chinese enterprises to move into other sectors. Family loyalty and the traditional mutual trust and sense of personal obligation among business associates make the necessary diversification moves more practical than they might have been otherwise. Flexibility and mobility are enhanced by ready access to working capital and financing for other purposes. Hence banking and other financial institutions in which Chinese enterprises had long excelled have expanded in recent years. This fact has in turn contributed to the expansion of Chinese-owned businesses in foreign trade and manufacturing. The owners of a number of Chinese enterprises stated in interviews in 1982 that their reputation of reliability as suppliers, backed by both trade and general credit, have given them a competitive edge against indigenous competitors, even among native clients.

##### B. *Locational Diversification and Internationalization*

When discriminatory treatment turns into violence, (as in Malaysia in May 1969 and Indonesia in 1974), the one-time immigrants could again emigrate. However, psychologically, the erstwhile alien entrepreneurs are no longer footloose transients. Hence a less drastic step is to move a part of one's business operations to less hostile jurisdictions. A

common practice is to set up affiliates, not necessarily in the same line of business, in foreign countries and then return as foreign investors enjoying special advantages. In the past, Singapore and Hong Kong have been on the receiving end of capital and entrepreneurial inflows of this nature, especially from Indonesia and Malaysia. Locational diversification has led to internationalization and the gradual emergence of Chinese multinationals. Chinese-owned international banking has successfully linked up with Western interests, mainly through Hong Kong and Singapore (Wu and Wu). The development of offshore banking in key vehicle currencies at both centers has contributed to the emergence of a new Chinese element in the international capital market.

The discriminatory policies of Indonesia and Malaysia give a large discretionary role to individual administrators and create a multiplicity of regulatory agencies in rule making, with or without advance notice. Besides, the public officials' salary scales are low. Not surprisingly, payment to officials for the revision or a more favorable interpretation of certain regulations so as to expedite matters has become such common practice that it could be virtually regarded as a normal business cost. At the same time, Chinese entrepreneurs must devote a disproportionate amount of energy to dealing with the bureaucracy. Where the native authorities do not regard the ethnic Chinese, realistically speaking, as potential competitors for political power, they are less reluctant to accept personal advantages from the Chinese. The potential damage to officials and other public figures caught in receiving favors from native businessmen would be immeasurably greater because the latter could be connected with the recipients' political opponents. This strange twist is known to have actually made the excessive regulatory practices work against the relative economic interests of the indigenous sector which it is the intent of some regulations to benefit. In those instances where native front men are installed in the executive suite as a token of indigenization, what emerges is a new rentier class, not a new crop of native entrepreneurs.

#### V. Chinese Entrepreneurship in an Orderly Society

What if the political environment were actually favorable? Such a case is presented by Singapore where profit making is protected by law and corruption is severely punished. First, such traditional characteristics as emphasis on personal relations and mutual trust in verbal contracts can still be detected but are becoming less prominent. Employment of family members in one's own or affiliated businesses is declining partly because many new employment opportunities have emerged. Clan associations and speech groups likewise have become less important because their previous welfare functions have been preempted by the government. Second, the trend is to supplant traditional and intuitive management methods with techniques learned in Western business schools, such as computerized information management, system analysis and return optimization, especially by younger businessmen. Third, because of Singapore's intimate involvement in the affairs of Chinese entrepreneurs based in Indonesia and Malaysia, the latter's style still has a residual impact on Singapore. As a result, Singapore's Chinese entrepreneurs appear to be only gradually, albeit unmistakably, acquiring the image of Western entrepreneurs everywhere. In time one may no longer be able to point to *Chinese* entrepreneurs as such in Singapore unless this "Westernization" is slowed down by the emergence of a new Singaporean business culture.

In conclusion, Chinese entrepreneurs in Southeast Asia now cover a wide spectrum, ranging from alien entrepreneurs in an *LDC* environment as postulated by some development theorists to entrepreneurs of increasing sophistication in a rapidly developing economy. Accordingly their entrepreneurial functions also vary widely, from devoting progressively more attention to improving productive and management methods to assigning top priority to pacifying the authorities. The self-perception of these entrepreneurs varies with others' perception of them; some are evolving from alien entrepreneurs

into simply entrepreneurs. It has proved to be difficult to replace them, repeated efforts to such an end notwithstanding. Nor can one dispute the lasting impact they already have left on the economies of Southeast Asia.

#### REFERENCES

- Fukuda, Shozo, *Kakyo Keizei Ron* [A Treatise on the Overseas Chinese Economy], Tokyo: Ganshodo, 1940.
- Hagen, Everett E., *The Economics of Development*, Homewood: Richard D. Irwin, Inc., 1975.
- Hoselitz, Bert F., *Sociological Aspects of Economic Growth*, New York: The Free Press, 1960.
- Kilby, Peter, *Entrepreneurship and Economic Development*, New York: The Free Press, 1971.
- Kunkel, John H., *Society and Economic Growth: A Behavioral Perspective of Social Change*, New York: Oxford University Press, 1970.
- McClelland, D. C., *The Achieving Society*, Princeton: D. Van Nostrand Co. Inc., 1961.
- Wu, Yuan-li and Wu, Chun-hsi, *Economic Development in Southeast Asia: The Chinese Dimension*, Stanford: Hoover Institution Press, 1980.

# The Levantines in Latin America

By WILLIAM GLADE\*

That alien entrepreneurs have historically figured prominently in the business life of less developed areas comes as no surprise. Nor is their success necessarily difficult to explain. Bearers of a dynamic industrial and business tradition, European and North American expatriates brought with them a number of specially valuable assets; for example, direct knowledge of and contacts in export product markets, familiarity with new production and organizational technologies, and access to superior capital supply sources. More problematic is the case of immigrants whose homelands were at least as traditional, preindustrial, and underdeveloped as the settings in which they attained entrepreneurial distinction. The "human capital" explanation that helps account for the success of errant British, American, German, and French business developers is obviously not germane to understanding the economic achievements of Chinese in Southeast Asia, Indians in East Africa, or Levantines in Latin America and West Africa. In the Latin American case, the Levantines, who began to arrive not long before the turn of the century, emerged as organizers of new firms in manufacturing relatively early in the industrialization process—well before extensive government intervention defined industrial policy. What circumstances gave these migrants such a comparative advantage in entrepreneurial endeavor that they came to play a role disproportionate to that of the nationals of the lands in which they settled?

## I. Analytical Framework

Over the years, the theoretically elusive concept of the entrepreneur has been char-

acterized differently by various scholars. In a common view, the function performed is that of coordinating other factors of production. For Knight, the pivotal task was to bear uncertainty; for Schumpeter, innovation, in any of several senses. The Harvard project in entrepreneurial history turned up still other conceptualizations (see Arthur Cole). In Latin America, whence comes the evidence considered here, the denotation of the term *empresario* is closest to the first of the foregoing, usually with some connotation that management of a new venture is involved.

The very range of functions that have been associated with entrepreneurship serves as an indication that the variable is probably a composite, one comprising a number of roles, attributes, and skills, the most appropriate mix of which varies with the situation. History is anything but prosaic regarding the possibilities in this respect. The assortment of talents that have yielded expansive management, a least common denominator of the various definitions of entrepreneurship, has been remarkably varied from setting to setting, so that it is not necessarily useful to deconstruct the composite variable except to recognize historical specificity.

Using an approach akin to demand and supply analysis, with some findings from sociology and anthropology, I shall review the experience of Levantine immigrants in Latin America to demonstrate how historical conditions conferred on them a comparative advantage in the supply of entrepreneurial resources. The differential supply of entrepreneurs from the Levantine community is examined as a historical product, just as, say, the aggregate supply of labor or of land in an economy is also historically generated. As will be seen, a structural explanation would appear to account adequately for the behavior in question; there is no need to posit such attitudinal or motivational variables as David McClelland and Everett

\*The University of Texas. I gratefully acknowledge support from the Latin American programs of the universities of Wisconsin, Texas, and California—Los Angeles at various stages in this project, and the most able assistance of Mary Elizabeth Wilkie.

Hagen have in their studies of entrepreneurial accomplishment.

## II. Demand and Pull Factors

Given that the presence of Levantine expatriates in Latin America involved international migration, the factors generating a demand for entrepreneurial behavior are essentially the same as those that pulled labor and capital to the region. Export-led growth made the last quarter of the nineteenth and the first quarter of the present century a time of exceptional expansion, and provided the means for opening up both new regions and new industries. Connected with this and the concomitant impetus to a widespread commercialization of local economic relations, there was considerable acceleration in the tempo of residentiary economic activity and a growth in its complexity. Foreign capital and technique, domestic capital, and, in the Southern Cone plus southern Brazil, immigrant labor in substantial numbers interacted to generate new patterns of internal trade and production, and hence a new array of business opportunities, alongside the external sector. Owing to conditions which kept wages low in national labor markets, other portions of Latin America failed to attract large-scale migration, but since they, too, enjoyed export expansion, smaller numbers of immigrants showed up there as well, coming as businessmen, engineers, and other specialists.

The quality of the statistical record for most of Latin America leaves much to be desired in respect of net Levantine immigration, since in some regions many immigrants also re-emigrated and comprehensive breakdowns by country of origin are not always available. Similarly, census information for earlier decades sometimes fails to report residents by foreign origin, while the American-born offspring of immigrants and those of their descendants who retained some measure of ethnic identification have, of course, not been enumerated separately. For these and other reasons, one cannot say for certain just how many Levantines, culturally defined, lived in Latin America at any particular point

in the past. Neither is there anything other than very crude estimates of the size of the resident ethnic community today.

Two things, however, do seem clear. First, the total number who came was quite small, certainly far behind the Spanish, Portuguese, Italians, Germans, and others. Second, much of the evidence suggests that Brazil, Argentina, Chile, Mexico, in roughly that order, received the most immigrants from the Levant and that other preferred destinations in the region included Cuba, Venezuela, Uruguay, and Colombia. It is believed that of those who went to Mexico and Cuba, some did so to facilitate an eventual entry into the United States.

Inasmuch as the foregoing were also the Latin American areas that participated most in the global economy of that day, the Levantine migrants were not evenly or randomly dispersed throughout Latin America. On the contrary, they clustered, reasonably enough, exactly where local economies were growing. This in itself, then, put them in particular proximity to the structure of business opportunities, that is, close to the demand for entrepreneurship (and labor) compared with the Latin American populations as a whole. For that matter, even within the receiving countries, immigrants tended to congregate in the more dynamic regions.

Examined in the light of what is known of the early immigrants, the proximity factor on the demand side becomes potentially even more significant. Notwithstanding the rural background of the majority of Levantine immigrants, few went to work in Latin America as farmers or farm laborers (the largest portion, in that day, of the total labor force). Neither do they seem to have gone in significantly numbers to work in construction, mining, railways, public utilities, government bureaucracy, or the liberal professions, to name some of the more dynamic labor markets. If we can construe the situation then prevailing as one of complementary macro-structural demands for labor and entrepreneurship, attention is then directed towards factors that tended to skew the attraction for Levantines more towards the latter. This, in turn, leads into the somewhat more

complicated supply-side conditions that caused the immigrants to be overrepresented in appropriating the new entrepreneurial opportunities.

### III. Supply and Push Factors

The pull factors for migration, which operated internationally and affected the labor markets of many countries, were joined in the case of the Levant by powerful push factors. Population growth pressed against the supply of agricultural land and exacerbated customary tension between Moslems and Druse, on the one hand, and communities of Christians on the other. The Suez Canal's opening displaced an overland trade that passed through the region en route to southern European ports and brought ruin to an export-oriented Levantine silk industry. In short, several centuries of growing commercial interaction with Europe, dominated by the French but involving others as well, ran into appreciable dislocation. Though for the most part rural in origin, the emigrants being pushed out were thus villagers who had lived in continuing contact with the active commercial life of the urban Levant, including its trade with France and elsewhere. Meanwhile, just prior to the earliest departures for Latin America, Lebanese had already begun to leave their homeland for Egypt and North Africa, where they were caught up in the quickening commercial ties of those places with Europe and more distant parts of the British and French empires.

Cultural (hence informational) ties with the West, including the United States, also began to multiply as Protestant missionaries arrived in the early nineteenth century to join the French and other Catholic clergy who had been working on the Latinization of the Levant—and of the Lebanon in particular—for several centuries. Both groups, as did the Russian Orthodox, sponsored schools for their co-religionists in the area, and additional Levantines, headed for the United States, joined the small number who, especially among the clergy, had long gone to Europe for advanced study. There was even

a slow build-up in the influx of foreign tourists, including two much publicized visits by a Brazilian emperor.

Significantly, in the peculiar political and cultural regime of the Ottomans, religious affiliation was foremost among the diacritica for differentiating groups and a major focus of loyalty and ethnic identity. Political conditions turned more oppressive under late Ottoman rule and since religious discrimination had already in the 1860's led to massacres of Christians amid a civil war, the option to emigrate became increasingly attractive, especially when, in the early 1900's, the Ottomans reversed previous policy and subjected Christian communities to military conscription. While we do not have complete information on how the characteristics of emigrants compared with the sending populations as a whole, it is established that they were overwhelmingly Christian (primarily Maronite, Melkite, and Greek Orthodox, with some Protestants to Brazil), and that the more Westernized even Francophile Lebanese were the most numerous, followed by Syrians and Palestinians.

The whole process of emigration was abetted considerably by publicity from shipping companies and their agents and by Latin American government-sponsored and private land-settlement and labor-contracting programs that advertised and operated in Europe, especially the Mediterranean countries. To an important extent, the semi-organized and quite substantial flows of people from northern Mediterranean ports to Latin America generated externalities (information, lower passenger fares, in-transit portside lodging, and the like) for the Levantines and other would-be emigrants in the region.

The informal but nevertheless extensive information system composed by the foregoing elements was extended further as Levantine families pooled resources to send members abroad to reconnoiter possibilities in the more promising Latin American destinations. These "scouts," in turn, sent and brought back accounts of opportunities and problems, triggering subsequent waves of emigrants who tended to cluster by village

and kinship linkages. As later emigrants left for America, additional information of a comparative sort became available in such ports of transshipment as Marseilles and Genoa, wherein out-bound travellers often spent upwards of several weeks awaiting connecting passage and comparing notes with other emigrants, as well as with returning migrants who furnished news on the latest conditions. Besides this not-inconsiderable pool of information, two other aspects of the migration process need to be remarked. It was not uncommon for immigrants to move on if their initial choice of location failed to live up to expectations, giving the Levantines a decidedly higher level of mobility than was typical of Latin American nationals as a whole. Further, immigrants often kept in at least intermittent touch with kin and friends who settled elsewhere, establishing an ethnically based network of contacts that was more extensive than was customary for the more parochial Latin American social and family structures of that day.

Lastly, the mode of the immigrants' usual incorporation into Latin American economic systems afforded them the additional advantages needed to account for their eventual entrepreneurial prominence. Owing to a virtual absence of entry barriers and to the fact that they could occasionally bring an initial inventory with them, a large portion of the early arrivals, including some women as well as men, took up peddling, in which pursuit capital was kept both liquid and mobile. Thereby the newcomer became quickly acquainted with trends and conditions in an evolving national market that was, typically, shot through with imperfections, not the least of which were on the market knowledge side. Almost invariably peddling accumulated the wherewithal for setting up shopkeeping in some attractive location, and in no few instances retailing led to wholesaling. Intra-ethnic networks of business contacts were formed through this branching and were in fact characteristic of the community. Later arrivals, therefore, had the benefit of experienced business counsel and apprenticeship in an occupation—mercantile activity—that more than any

other could realize the potentialities of the market as an information system.

Although Levantines lacked the rotating credit associations that Chinese and Japanese emigrants took with them, family and kinship networks helped to mobilize loan and equity capital. Riding the crest of an expanding commercialization of economic life, those in business were, moreover, advantageously positioned for further capital formation—and, for structural reasons as well (namely, ties with a growing distributional apparatus), for an eventual move into manufacturing. Burton Benedict and Samir Khalaf and Emilie Shwayri have detailed the important social resources that could be activated to good account by family-based entrepreneurs, while Paul Siu has called attention to the economic implications of the sojourner status held by the early cohorts of immigrants. Obligated to repay family- and kin-organized advances and expecting eventually to return with an investible surplus to use in their homelands, most came with a built-in high marginal propensity to save, and a disposition to invest those savings in one of the relatively liquid business opportunities their multiple information networks might turn up. (That most nevertheless stayed on and, in fact, sought rapid assimilation enlarged the community's investment funds through time.) Whether this situation led to a higher household savings rate than prevailed on average among the local population cannot be established with certainty, though a reasonable presumption is that it did. What seems clearer is that on the investment side of the equation, Levantine immigrants had, thanks to their social cohesiveness and informal community organization as well as to their occupational specialities, access to a range of alternatives quite a bit richer than was ordinarily in view for most of the local populations.

In conclusion, the experience of the Levantines, some of whom eventually went on to establish extensive business empires, helps validate the standard emphasis on factor mobility as a condition for development. Still more, it speaks to the crucial importance of information, even, as the World Bank noted

in its country report on Nicaragua, in the initial phases of development. Not often does one find the strong empirical confirmation of this latter theoretical point that the experience of these minority entrepreneurs would seem to supply.

#### REFERENCES

- Benedict, Burton, "Family Firms and Economic Development," *Southwestern Journal of Anthropology*, Spring 1968, 24, 1-19.
- Cole, Arthur H., *Change and the Entrepreneur: Postulates and Patterns for Entrepreneurial History*, Cambridge: Harvard University Press, 1949.
- Hagen, Everett E., *On the Theory of Social Change*, Homewood: Dorsey Press, 1962.
- Khalaf, Samir and Emilie Shwayri, "Family Firms and Industrial Development: The Lebanese Case," *Economic Development and Cultural Change*, October 1966, 15, 59-69.
- McClelland, David C., *The Achieving Society*, Princeton: D. Van Nostrand, 1961.
- Siu, Paul C. P., "The Sojourner," *American Journal of Sociology*, July 1952, 58, 34-44.

## WOMEN AND HEALTH

### Women and Absenteeism: Health or Economics?

By LYNN PARINGER\*

Full-time workers in the United States lost 3.5 percent of scheduled work time due to unscheduled work absences in 1978. Nearly 7 out of 100 workers experience at least one spell of absence a week. However, U.S. absentee rates have remained fairly constant over the past several years and are generally well below the rates of Western Europe (for example, West Germany lost 11 percent of scheduled work time to unscheduled work absences, Holland lost 10 percent, and Italy lost 15 percent in 1972-73). Part of the higher absentee rates in Western Europe is thought to be caused by the high prevalence of sick leave availability in these countries.

Work loss from unscheduled absences is substantially higher among female workers than among male workers in the United States. In May 1978, women lost 4.3 percent of scheduled work time compared to 3.1 percent for men. Absence incidence rates are also higher for women; 8.6 percent of women workers experience a spell of absence during a typical work week compared to 5.4 percent of men workers. The large majority of unscheduled work absences are due to illness and injury. The average female worker lost 5.4 days from work for health reasons compared to 4.7 days for the average male worker in 1979.

There have been several recent research efforts on the causes and consequences of worker absenteeism. Steve Allen (1981) found evidence that observed differences in absentee rates between the sexes could be accounted for by differences in marginal earnings and the flexibility of the work schedule. My 1978 study found that women generally lost fewer days from work than men for most common illnesses. There has

been little research to date on gender differences in work loss due to illness that takes into account differences in the occupational distribution of workers, and differences in marginal earnings as well as variations in health status. This paper reports on an empirical investigation of the determinants of gender differences in illness related absentee rates. In this investigation, labor market characteristics such as differences in earnings and occupational mix between men and women, differences in health status, and the impact of family are analyzed using a regression framework.

#### I. Conceptual Framework

Unscheduled work absences impose economic costs on employers and cause uncertainty with respect to the size of the work force on any given day. One response of employers may be to hire more workers, thereby increasing production costs. Another response is to penalize workers with high absenteeism by paying them less or promoting them less frequently in jobs that are key to the production process. Additionally, firms may screen out workers who are expected to pose absentee problems.

Depending on the nature of the job, employee absenteeism may impose different costs on employers. In occupations characterized by general training and where employees are easily replaced, firms may be less concerned with absence rates. Also, where tasks can be accumulated and deferred to a later date, the cost of worker absences will be lower. Firms hiring employees in these occupations will likely have less concern over worker absences and may offer flexibility in hours to accommodate the workers. By contrast, in occupations where specific training predominates and the employee is an in-

\*Associate professor of economics, California State University-Hayward.

tegral part of the production process, absenteeism is less likely to be tolerated and prospective employees may be more closely screened by employers. The result is a barrier to entry for workers perceived to be less dependable.

From the worker's perspective, the decision to work while ill can be analyzed within the traditional market-work/home production context. Individuals are assumed to maximize utility. Each individual must allocate a fixed block of time ( $T$ ) to work ( $T_w$ ) and consumption ( $T_c$ ) activities. The arguments in the utility function are income ( $I$ ), time at work ( $T_w$ ), and time spent in consumption activities ( $T_c$ ):

$$(1) \quad U = U(I, T_w, T_c).$$

The marginal utility of time spent at work ( $T_w$ ) can be either positive or negative depending on whether workers derive satisfaction from their job in addition to the utility derived from the resulting income. To the extent that absenteeism results in fewer promotions and lower wage increases, time spent at work in the current period may enter the utility function with a positive sign thereby capturing some dynamic aspects of the labor market. Time spent at work can be divided into well time ( $T_{ww}$ ) and sick time ( $T_{ws}$ ). Similarly, consumption time can be broken down into consumption time while well ( $T_{cw}$ ) and time while ill ( $T_{cs}$ ). For simplicity, I assume that  $T_{cs}$  represents time missed from scheduled market work activities. Other analyses have generally assumed that time absent from work for health reasons was completely lost time and therefore unavailable for any consumption activities (see Michael Grossman, 1972a). This model relaxes that assumption by allowing individuals to derive utility from both time spent in consumption activities while well and sick. The modified utility function is thus

$$(2) \quad U = U(I, T_{ww}, T_{ws}, T_{cw}, T_{cs}).$$

The income available to the individual is comprised of unearned income and wage income for time spent at work ( $V + W(T_{ww} + T_{ws})$ ). In addition, if workers are eligible

for sick leave or other wage reimbursement schemes for time absent, they receive an additional  $WS(T_{cs})$  where  $S$  represents the fraction of wages which are reimbursed for sick time and  $W$  = wage rate. Substituting this into the model gives

$$(3) \quad U = U((V + W(T_{ww} + T_{ws}) + T_{cs}SW), \\ T_{ww} + T_{ws}, T_{cs}, T_{cw}).$$

Workers act to maximize their utility subject to the availability of time ( $T$ ), the wage rate ( $W$ ), and the availability of wage reimbursement ( $S$ ) for time lost from work due to illness. Workers weigh the marginal benefits of missing time from work against the marginal costs which may include lost earnings and reduced promotional opportunities. The impact of a change in the wage rate on absenteeism is ambiguous due to the income and substitution effects. In general, an increase in the wage reimbursement rate for time missed (an increase in  $S$ ) should increase the number of days missed from work. Factors that increase the marginal utility of time spent in home production should also increase the willingness of individuals to miss work. Within this framework, I generate a set of hypotheses regarding absentee differentials between men and women based on differences in economic factors, in health status, and in other variables which proxy the marginal costs and benefits of missing time from work.

Using mortality measures, a woman's health status is considerably better than that of a man. White females born in 1970 could expect to live 8 years longer than while males born in the same year. The higher life expectancies for women suggest that they may have the potential to function as productive members of society for a longer time period than their male counterparts. However, when one includes measures of morbidity in comparing the differences in health status between men and women, a somewhat different picture emerges. Women are more likely than men to report their health as being fair or poor, report more chronic conditions, and report more days of restricted-activity days

due to illness than men. Women also use significantly more medical care than men, even after controlling for obstetrical and gynecological care.<sup>1</sup> Part of the difference in reported restricted-activity days between the sexes may reflect greater investments in health on the part of women. Women may recognize and treat health problems at an earlier stage than men, thereby reducing the duration of the illness and possibly preventing the onset of more serious conditions. Evidence to this effect is the fact that, although women are more likely to miss a day of work than men for illness, the amount of time lost per absent worker is lower for women than it is for men. Also, when one compares work loss rates for men and women of the same age, health and marital status, the absentee differential between men and women falls with age (U.S. National Center for Health Statistics, 1974). This is particularly evident for married women in good or excellent health whose absence incidence falls below that of men in the aged 45–64 category. The two-week absence incidence rate for these women was 3.7 compared to 3.9 for men in the same category. In contrast for married women aged 25–44, the incidence rate was 5.4 compared to 4.6 for men.

To capture differences in health status between workers, the empirical model includes a perceived health status variable and the age of the individual. Both sets of variables are expected to be positively related to work loss due to illness.

Differences in earnings and sick leave benefits between men and women could also account for variations in absentee rates. In jobs where sick leave is unavailable, the short-run cost to workers of missing a day of work is the loss in earnings for the time missed. Lower earnings for women reflect lower costs of work loss. The potential impact of the earnings gap is ambiguous from a theoretical standpoint, since changes in earning pose both an income and a substitution

effect. When women are not the only working member of the family, one might expect the substitution effect to dominate.

In addition to earnings, the availability of wage reimbursement schemes for work loss due to illness may affect the relative absence rates for men and women. If women are more heavily concentrated in jobs where such benefits are prevalent, one would expect higher absentee rates among women.

One way of measuring the incremental earnings a worker would receive if he or she attended work while ill is to combine the hourly earnings and wage reimbursement rate for sick time into a variable that reflects the sick leave adjusted marginal wage. Such a measure can be calculated as  $(1 - S)W$  using the model. Because this variable may be endogenous in the model, a value was predicted for this netwage using an instrumental variable approach.

Another reason that work loss differential might exist between the sexes may be due to differences in the occupational distribution of men and women. In occupations characterized by primarily general training, absentee workers are easily replaced and job tasks may be easily accumulated. The cost to the firm of unscheduled absences may therefore be less and penalties imposed on absent workers may be smaller. One would expect absentee rates for both males and females to be higher in these occupations. Among white-collar employees, work loss rates among clerical workers are about 50 percent higher than work loss rates for managers. Since clerical jobs are dominated by women, while professional and technical jobs are dominated by men, occupation may be an important factor giving rise to the observed sex differential. Occupation is controlled for in this model by segmenting the sample according to occupation. Results from two different occupational categories are presented here; professional and managerial workers, and workers in clerical or sales occupations.

One hypothesis regarding the higher work loss rates for women concerns their traditional dual role as both home producers and labor market participants. When a woman becomes ill, both her home and market pro-

<sup>1</sup>Jody Sindelar (1982) presents an analysis of the gender differential in medical care utilization. David Mechanic (1978) discusses the reasons for reported morbidity differences by sex.

ductivity are affected. Because of her dual responsibilities, the full impact of ill health for the female may carry with it greater costs than the lost earnings associated with missing work. Thus, women may have a lower work loss threshold to a given illness than men because there is a greater payoff to the family if the woman responds to the illness early. Consistent with this notion is cross-sectional data which indicate that married women lose more hours from work for unscheduled absences than unmarried women. The presence of family responsibilities is measured by two variables; one indicating whether the individual is married with the spouse present, and another indicates the presence of other dependents in the household.

Education is included in the model because as, Grossman (1972b) has suggested and empirically found, education is positively related to efficiency in the production of health. More educated individuals should be better able to care for themselves and thus less inclined to miss time from work for illness related reasons.

The total family income is included in the model as a measure of the family's asset position. The assumption of diminishing marginal utility of income leads one to expect higher work loss among workers from high income families. However, to the extent that income is negatively correlated with health status this variable may pick up some of the health status determinants not captured in the model.

## II. Data and Empirical Results

In the empirical investigation, the 1974 *HIS* data was used because it includes some of the most current data on work loss, illness and sick leave compensation. It represents a random sample of approximately 120,000 members of the civilian non-institutional population. Data used in this study are unweighted and include white males and females below the age of 65 who were employed at the time of the survey. All work loss data pertains to the two week period prior to the interview.

Using hours lost from work as the dependent variable, regressions were estimated

separately for both men and women. The data were also split by occupation into two groups: professional/technical and clerical/sales. In one set of regressions, all workers were included; in the second, only those who reported missing some time from work for health reasons. Details of the regressions results are reported in my earlier study. Here I summarize the primary findings.

(i) The regressions indicate that perceived health status is an important predictor of hours lost when all workers are included in the regression. (ii) Age is significantly related to the number of days which absent workers miss from work for an illness. (iii) The size of the age coefficient is more than twice as great for men as women in both occupational categories. (iv) The sick leave adjusted incremental wage is negatively related to hours lost among male professional and technical workers. However, even among these workers it is not a significant determinant of work loss duration for those who missed some time from work. It is positively related to female work loss in the second set of equations although the magnitude is very small.

All else equal, married female professional and technical employees are less likely to miss work and their work loss duration is shorter than unmarried women. In all of the regressions, the presence of dependents is negatively related to work loss. Education and family income are generally not significant variables in either hours lost regression.

## III. Conclusion

The evidence reported here indicates that health status and age are principle determinants of work absences. The impact of health status on work absence is significant for males and females, and in both clerical/sales and professional/technical occupations. Poor health has a substantially larger impact on male work loss in professional/technical occupations than it does on female work loss.

Age is an important determinant of absentee duration for both sexes. The age coefficient on men is three times what it is for women professional and technical workers. The coefficient on age among clerical and

sales workers is twice as great for men as for women. This suggests that age may have a more deleterious effect on male productivity in the more demanding white collar jobs. By contrast, the effect of age on female work loss is smaller in professional/technical jobs. The differential impact of health status and age on work loss for men and women may in part result from gender differences in the early perception and treatment of illness. Previous research indicates that women are more likely to invest in their own health through greater use of preventive and other medical services. This may result in slower depreciation of their health stock as they age. This is consistent with larger declines in work loss rates and longer life expectancies for women compared to men.

Economic variables appear to have little impact on the amount of time workers miss from work. The implication of this is that previously stated notions regarding the impact of sick leave on worker absenteeism generally overstate the relationship. Marginal changes in the economic incentives facing workers are not likely to alter worker absentee rates.

The presence of family responsibilities appears to reduce the amount of time missed from work, particularly among women. Married women with families may exhibit a stronger labor force attachment than unmarried women. Because of their dual responsibilities, they may also invest more in their health thereby lowering their illness rates.

The results of this study add to our understanding of the gender differential in illness absentee rates. Within an occupation category, work loss rates for both male and

female workers are affected primarily by health status considerations and not by economic variables. Furthermore, age has a considerably larger impact on work loss for men than women, suggesting that perhaps women's greater use of medical care and early response to illness may have payoffs later in terms of lower absentee rates during middle age as well as longer life expectancies.

## REFERENCES

- Allen, Steve, "An Empirical Model of Work Attendance," *Review of Economics and Statistics*, February 1981, 63, 77-87.
- Grossman, Michael, (1972a) *The Demand for Health: A Theoretical and Empirical Investigation*, New York: Columbia University Press, 1972.
- \_\_\_\_\_, (1972b) "On the Concept of Health Capital and Demand for Health," *Journal of Political Economy*, March/April 1972, 80, 223-55.
- Mechanic, David, *Medical Sociology*, 2d ed., New York: Free Press, 1978.
- Paringer, Lynn, Determinants of Work Loss and Medical Care Utilization for Specific Illness," unpublished doctoral dissertation, University of Wisconsin, 1978.
- Sindelar, Jody L., "Differential Use of Medical Care by Sex," *Journal of Political Economy*, September/October 1982, 90, 1003-19.
- U.S. National Center for Health Statistics, *Current Estimates from the Health Interview Survey: United States, 1974*, Series 10, Number 100, U.S. Dept. of HEW, September 1975.

# Women and the Use of Health Services

By GAIL R. WILENSKY AND GAIL LEE CAFFERATA\*

It is well documented in most industrialized countries that women have higher rates of use of health services than men. The probability of seeing a physician, the number of visits per user, the probability of using hospital inpatient services, and the use of prescription drugs all are higher for adult women than adult men. (R. Andersen et al., 1976; L. Verbrugge, 1982). Although hospitalized men have a longer length of stay and there are some reversals in this pattern by age, men over 50 being more likely to be hospitalized than women, and boys using more of all types of health services than girls, there is no denying the overall pattern: Even after controlling for pregnancy-related services, women use more health services than men.

To answer the question why this is so, researchers have turned to models developed to explain the use of health services. In a well-known sociological model, Andersen et al. (1975) described three factors which affect the use of health services in general and by both men and women: personal or predisposing characteristics, resource availability or enabling characteristics, and the need for health care. More specific sociological models propose "sex role" theories which explain how society engenders differential perceptions of morbidity for men and women; "social support" theories which relate to integration into social support networks; and "stress" theories which focus on the reactions to perceived or differential levels of stress for men and women (see Cafferata et al., 1981).

Economists normally regard the demand for a good or service as dependent on prices (both own- and relative prices, time price,

and money price), income, and consumer preferences. In a well-known economic model specifically adapted to health, Michael Grossman (1972) posits utility-maximizing consumers who seek to maximize their stock of "health capital," and who demand medical care to guard against the depletion of health. Despite their origin in different disciplines, both the Andersen model and the economic models result in relatively similar predictive factors of determinants of medical care use.

The purpose of this paper is not to discuss the relative merits of these and other models, but to explore in some empirical detail the issue of differential use of health services by men and women. For this, we use a 1977 data set from the National Medical Care Expenditure Survey (*NMCES*). Detailed information was collected in *NMCES* for calendar year 1977 from a sample panel of 40,000 individuals chosen so as to be representative of the U.S. population (S. B. Cohen and W. D. Kalsbeek, 1981). Information was collected on the use of various types of medical services, expenditures, and sources of payment for each type of service, whether the patient or physician initiated each visit during the year, and the medical condition associated with each visit or disability day. Extensive economic and demographic data were collected as well (S. Bonham and L. S. Corder, 1981).

These data permit us to address three issues: one, were U.S. women in 1977 in fact more likely than men to exhibit the characteristics known to be associated with the use of health services and did they exhibit these characteristics at different levels? Second, were women more sensitive or responsive at the margin to the presence or absence of these factors than men? In other words, did women respond more directly to a day of disability or a dollar of out-of-pocket cost than men? Third, were these effects similar when health services were used for particular illnesses?

\*National Center for Health Services Research (NCHSR), DHHS. We are indebted to Pamela Farley and Judith Kasper for analytical advice, and to Lita Manuel for data processing. The views herein are our own, and no official endorsement by the NCHSR is intended, or should be inferred.

We will consider the use of ambulatory physician services. Other ongoing *NMCES* studies are examining the differential use of hospital services and prescription drugs (including psychotropic drugs).

### I. Some Factors Expected to Affect Use

With regard to our first question, economic theory predicts that demand will increase if the price of the service decreases. In the case of visits to the physician, the relevant price variable is the amount of the bill paid by the patient. Because of widespread insurance coverage, this out-of-pocket price is not the same as the gross money price per visit. In addition to the money price, the time spent in getting to the doctor, in waiting rooms, and during treatment also represents a cost that can be expected to reduce the number of visits. This time effect can be quantified by considering time spent in minutes or, alternatively, valuing the opportunity cost of time by multiplying time in minutes by the value of time for each individual. At various stages of the analysis, we have used both measures, but here present only time in minutes (Table 1). In fact, the supposedly lower time costs of women are one of the explanations frequently given as to why women use more health services than men. However, as Jody Sindelar (1982) points out, not only does the lower average wage paid to women not necessarily mean that their time has a lower value, but the effects of opportunity cost of time on the use of medical care cannot be predicted in the general case and must be determined empirically.

In addition to the economic variables, tastes, health status, age, and other demographic factors affect the demand for medical services. If the demand for visits reflects a demand for better health, we would expect the use of services to be a function of perceived health status and sick days (defined here as the unduplicated sum of days in bed, days lost from work or work around the house, and days when normal activities are curtailed). Apart from age, which is a proxy for health status and health care need, other sociodemographic variables which have been

TABLE 1—MEAN VALUES AND PERCENT OF PERSONS WITH SELECTED CHARACTERISTICS HAVING AT LEAST ONE AMBULATORY PHYSICIAN VISIT, 1977

	Male	Female
Age		
18-24	17	16
25-44	36	36
45-64	31	30
65 and over	16	18
	100	100
Race (percent)		
White	88	85
Black	8	10
Hispanic	3	4
Other	1	1
	100	100
Perceived Health		
Excellent	43	36
Good	38	43
Fair	13	16
Poor	5	5
	100	100
Education		
8 years or less	18	17
9-11 years	18	18
12 years	35	40
13-15 years	14	13
16+	15	10
	100	100
Work Experience		
Full year, > 35 hr/week	66	34
< 35 hr/week	6	12
Part year, > 35 hr/week	6	6
< 35 hr/week	4	7
Never worked	19	41
	100	100
Income		
Poor/Near Poor (up to 125 percent of poverty line)	11	17
Other Low Income (125 to 199 percent of poverty line)	13	14
Middle Income (2-4 times poverty line)	36	36
High Income (4 times + poverty line)	40	33
	100	100
Average Number of Sick Days (Mean)	14	14
Average Percent of Visits with a Chronic Condition	28	24
Average \$ Spent Out-of-Pocket per Visit	14	17
Average Waiting Time per Visit (Minutes)	52	55

Source: National Medical Care Expenditure Survey, NCHSR.

shown to affect the use of health care are race or ethnicity, education of the family head, employment status, and family structure. The latter two are of particular interest because they should reflect flexibility in scheduling and constraints on time as well as possible social support relationships. Stress variables such as an unemployed spouse, ill child, death in the family, or change in the family structure have been examined for their differential effect on men and women, but in our analyses showed a consistent lack of effect.

In fact, as far as most of the characteristics associated with the use of services are concerned, our data showed not much of a difference between men and women. There was no difference in race, age distributions, or in the average number of sick days. There are several variables in which they differed a little. A smaller share of visits by females were associated with chronic conditions, their average out-of-pocket expenditure was a little higher, and their average waiting time a little longer. There are a few variables in which the differences were larger: women were less likely to be in excellent health and more likely to be in good or fair health; they were more likely to have finished high school but less likely to have finished college; they were much less likely to have worked full time all year, and more likely to have never worked; and they were more likely to be in poverty and less likely to have had high family incomes.

The second issue we address is whether the factors which are important in explaining health care use differ in their effects for men and women. To test this, we include all of the variables predicted to affect use, but also include intercept and slope dummy variables for women. The coefficients of the interaction terms with sex tells us whether the coefficient for a particular variable is different for women and also the direction of the effect. The dependent variable in each of the equations is the *log* of the number of ambulatory physician visits. (The dependent variable is logged because of the skewed distribution of physician visits and expenditures which is approximated by a lognormal distribution.)

The third and probably most interesting issue is whether these differential effects, if any, persist in the case of visits for particular conditions. To test this, we estimated equations predicting the use of physician services for an acute condition, colds/influenza, and for the chronic condition of diabetes.

The results of these estimations are summarized in Table 2 in terms of positive and negative effects of the statistically significant variables on the number of physician visits. The first estimate is for all adults, the second for the working population, the third for visits for colds, and the fourth for visits for diabetes. (The corresponding regression coefficients and the *t*-statistics can be obtained from the authors.) The results shown in Table 2 reflect the behavior of adults with at least one physician visit. Equations predicting the probability of a visit and expenditures for visits also can be obtained from the authors. The factors explaining the use of physician services from the *NMCES* data are discussed only briefly here as they have been described in detail elsewhere (Wilensky et al., 1981).

As expected, for all adults and working adults, we find the relationships between money price and use negative and significant. Waiting time is also consistently negative and significant. The remaining independent variables either have the expected sign or are insignificant. Also as expected, the strongest and most consistent factors predicting use are the various measures reflecting "need" for services—perceived health status, the number of sick days, and the share of visits with a chronic condition. This suggests that perceived need and not economic factors are the most important determinants of health care use, which is not very surprising.

When the analysis is limited to visits for a particular condition, economic variables become even less significant and the medical need variables more significant. While the effect of money price remains negative, it is significant only for diabetes; time is not significant. Most of the variance is explained by perceived health, sick days, the presence of a chronic condition, and the shift term for sex.

TABLE 2—FACTORS AFFECTING THE USE OF PHYSICIAN SERVICES: SUMMARY OF STATISTICALLY SIGNIFICANT MULTIPLE REGRESSION VARIABLES WITH POSITIVE OR NEGATIVE EFFECT ON NUMBER OF VISITS

	Variables with Positive Effect on Use		Variables with Negative Effect on Use	
	Without Interaction	Interaction with Female Sex	Without Interaction	Interaction with Female Sex
All adults <sup>a</sup>	Age No. of sick days Chronic condition Not in excellent health Education Part-year employment Physician density	Sex No. of sick days Chronic condition	Net price Visit time Being black	Net price
Working Population <sup>b</sup>	No. of sick days Chronic condition Not in excellent health Physician density	Chronic condition Education	Net price Visit time	
Visits for Colds <sup>c</sup>	No. of sick days Chronic condition Not in excellent health	Sex	Nontraditional family	
Visits for Diabetes <sup>d</sup>	Being black Age No. of sick days Physician density		Net	Physician density

Source: National Medical Care Expenditure Survey, 1977, NCHSR.

<sup>a</sup> $R^2 = .20$ ; joint  $F$  test/all female variables = 20.94.

<sup>b</sup> $R^2 = .21$ ; joint  $F$  test/all female variables = 12.01.

<sup>c</sup> $R^2 = .30$ ; joint  $F$  test/all female variables = 3.52.

<sup>d</sup> $R^2 = .24$ ; joint  $F$  test/all female variables = 1.38.

## II. Do Women Respond Differently to Factors Affecting Use?

Using sex as a categorical variable allows for a shift in the intercept term; the interaction of being female with each of the other variables allows the slope of each of these variables to change for females; the joint  $F$  test shows whether females differ from males including both slope and intercept terms.

Being female obviously makes a difference in the use of physician services. The joint  $F$  test is significant at the .01 significance level for all equations other than diabetes; the latter is significant at the .05 significance level. Table 2 shows that female sex, *ceteris paribus*, is a positive predictor of use by all adults and for visits for colds and influenza; it is not a predictor for the working population or for diabetes, however. What we did

not foresee is that females do not seem to differ from men in their response to economic variables, although there are some exceptions. In the all-adult estimation, females are more responsive to the money price than males. Having more education also increases the use of physician services for working women. The most persistent differences occur in the response to general medical need variables. In terms of variance in the overall use of physician services, women are more responsive to chronic conditions than men, and this is true for all adult women and working women as well.

## III. Discussion

The question we addressed in this paper is whether it is possible to establish specific factors which explain some of the higher use

of health services—in this case, physician services—by women. Is their use higher because their time and time price, their preferences, or their need for medical services differ, or is it that they are more sensitive at the margin to one or more of these factors? Or are these factors acting in conjunction?

On the one hand, women have slightly higher out-of-pocket expenses per physician visit and wait slightly longer. They also either react like men or are even more sensitive than men to time and money costs. Also, when time is valued rather than measured in minutes, women are more responsive to time costs than men. These effects, however, would suggest less use of services, not more.

The major factors associated with higher use by women are a lower perceived health status and a greater responsiveness at the margin to chronic conditions, although this is not true for disability days. In addition, the fact of being female is important for adult women in general and for some specific illnesses, but not for working women. Does this mean that women have illnesses that are more responsive to medical treatment? Are they more likely to seek care given medical symptoms, or more likely to have visits initiated for them by physicians? Although it is difficult to be sure, it appears as if all of the above may be true to some extent.

In an extensive review of survey and epidemiological data, Verbrugge concludes that women have different illness patterns and it appears they are sicker (i.e., they report more acute and chronic conditions), although it is difficult to eliminate behavioral responses and potential reporting bias from these measurements. Women do appear to respond more to illness, that is, they have more disability days per acute condition and also higher rates of restricted activity or bed disability for chronic conditions. This may be due to the fact that women have different conditions from men (which they do), or that these conditions are more severe. Perhaps women have differential perceptions of the efficacy of medical care, or we may, in part, see the influence of other social factors. If the latter is true, however, it reflects something we have been unable to measure.

On the supply side, it does not appear that physician initiation of visits is a major factor in explaining the differences in use patterns. The demand equations were reestimated with differences in demand decomposed into whether it is attributable to the patient or the physician. The differential responsiveness of females to perceived lower health status and the presence of chronic conditions was found to be attributable to an increase in patient initiated, not physician-initiated visits. The overall shift variable associated with sex, however, was about equally attributable to the patient and to the physician, implying that about half of this difference is because physicians initiate more visits for women than they do for men but, in addition, women also initiate more visits for themselves than do men.

The conclusion of our data set seen in conjunction with other data are the following: women are sicker or at least report more conditions; women respond somewhat differently to the factors that influence use, but the differences are not usually large. More interestingly, they are not always in the direction that would suggest greater use. For illness-specific visits, men and women respond more similarly than for all visits; and as more women join or remain in the labor force, and as their time costs increase, the differential use of physician services between men and women may well decline.

## REFERENCES

- Andersen, R. et al., *Equity in Health Services*, Cambridge: Ballinger, 1975.
- \_\_\_\_\_, *Two Decades of Health Services*, Cambridge: Ballinger, 1976.
- Bonham, S. and Corder, L. S., *National Medical Care Expenditure Survey: Household Interview Instruments*, National Center for Health Services Research, National Health Care Expenditures Study, Instruments and Procedures Series 1, DHHS Publication No. (PHS) 813280, Hyattsville, 1981.
- Cafferata, Gail et al., "Family Roles, Structure, and Stresses and Sex Differences in Obtaining Psychotropic Drugs" presented at the American Public Health Association

- meetings, Los Angeles, November 1981.
- Cohen, S. B. and Kalsbeek, W. D., *National Medical Care Expenditure Survey: Estimation and Sampling Variances in the Household Survey*, National Center for Health Services Research, National Health Care Expenditures Study, Instruments and Procedures Series 2, DHHS Publication No. (PHS) 813281, Hyattsville, 1981.
- Grossman, Michael, *The Demand for Health: A Theoretical and Empirical Investigation*, New York: Columbia University Press, 1972.
- Sindelar, Jody, "Differential Use of Medical Care by Sex," *Journal of Political Economy*, October 1982, 90, 1003-19.
- Verbrugge, L., "Sex Differentials in Health," *Public Health Reports*, September-October 1982, 97, 417-37.
- Wilensky, Gail et al., "The Role of Time and Money in the Demand for Medical Care," paper presented at the American Economic Association meetings, Washington, December 1981.

# Time Allocation, Market Work, and Changes in Female Health

By BARBARA WOLFE AND ROBERT HAVEMAN\*

The past two decades have seen remarkable time reallocations of working-age women among child care, market work, housework, and leisure activities; indeed, the labor force participation rate of women aged 25–34 increased from 36 to 65 over this period. These reallocations have sparked speculation regarding their effects on female health status. One view suggests that increased female market work will contribute to health status, via its contributions to self-esteem and social contacts and support. A second view emphasizes the mortality differences between men and women, and suggests that female morbidity and mortality patterns will resemble those of men as the two sexes become increasingly similar in their exposure to the hazards and stresses of market work. A third view notes that the increasing market work of women is not associated with equivalent decreased time spent in child care and housework, implying an increase in the total demands on the time of women who work. It is suggested that the resulting decrease (increase) in time allocated to leisure (demanding and stressful work activities) will have deleterious effects on women's health status.

For several reasons, it is difficult to sort out the determinants of changes in health status over a period of time. Clearly such changes depend on the age of the individual and especially her health at the beginning of the period over which measurement occurs. In addition, the individual's access to and utilization of health care and other resources during the period are likely determinants. And, while activities engaged in during a period (for example, hours of market work, work in hazardous or demanding circum-

stances, demands of "dual roles") are likely to have an effect on changes in health status over the period, the extent of these activities is likely to be endogenous, depending on initial health status. Data limitations add to the difficulty of obtaining reliable estimates of the determinants of health status changes. Most basically, longitudinal data appear necessary to identify the determinants of changes in health status or the role of activities on health at a point in time. While certain activities and habits (for example, exercise, diet, cigarette and alcohol use) are likely to affect health, information on these variables is often not available in micro data. Detailed time allocations among child care, housework, and leisure activities are typically not available, and even if available, the composition of activities within each class are unknown. Finally, the inadequacy of reliable health status indicators plagues this question at least as much as any in the health economics area.

In this paper, we first briefly relate what is known about the determinants of women's health and the impact of time allocations on changes in health. We then describe our conceptual model for analyzing the effect of time allocation on changes in health status. Finally, we present our empirical estimates of the relevant parameters of this model, based on panel data and techniques to deal with initial and changing health heterogeneity.

## I. Time Allocation and Women's Health

The economics literature on the determinants of women's health in the context of time allocation is sparse. Three lines of work relate to the issues addressed in this study. The first, work on the supply and demand for health, is based on the framework proposed by Michael Grossman (1972). In this model, health status depends upon initial health, the marginal costs and benefits of health determining investment, and the exogenous effect of health depreciation related to age and

\*Associate professor, department of economics and preventive medicine, and professor, department of economics, respectively, University of Wisconsin–Madison. The contribution of Mary Kay Plantes is acknowledged, as is the assistance of Louise Cunliffe, Nancy Williamson, and Michael Przybylski.

hazardous and stressful activities. The higher the wage rate of the individual, the greater the demand for good health. Conversely, high wage rates raise the cost of own-time health investments and encourage work time, which is taken to increase the health depreciation factor. Hence, the net effect of high wage rates on health is uncertain, although at high levels of aggregate time demands, less time will be spent producing health and thus the depreciation effect is likely to dominate.

The second set of contributions concerns the effect of health on labor force participation (for a recent discussion see Anne Bartel and Paul Taubman, 1979), and suggest that poor health in period  $t - 1$  reduces both labor supply and wage rates in period  $t$ , with larger effects found with health in  $t - 1$  being endogenous. This result emphasizes the complex relationship between health, work, and wage rates, and indicates the need to carefully distinguish the independent contributions of work (if healthy) and early health in explaining later health.

The third line of work concerns the patterns of time allocation by women (C. R. Hill and F. P. Stafford, 1974, 1980; A. Leibowitz, 1975). This work suggests that both the time allocated to market work and the time spent caring for small children are positively related to education, and that there is relatively little substitution among hours spent in market work, child care, and housework. Increased market work, hence, tends to result in nearly an equivalent increase in total time demands. Because increases in total time demands come at a cost of personal-care time and leisure losses, it is important that such demands be explicitly considered as determinants of health status changes.

Sociological and epidemiological literature (for example, I. Waldron, 1978, 1980; L. M. Verbrugge, 1980) has also addressed the determinants of women's health. Certain aspects of market work (hazardous, time pressured, or repetitive activities, the tendency to take fewer rest days when ill, and the contribution of market work to total time demands) are found to be negatively related to health; while the social support and independence aspects of market work tend to contribute to health. The epidemiology liter-

ature on the influences of work on women's health emphasizes the role of particular hazards (for example, cotton dust, carcinogens) which contribute to specific health problems. Much of the discussion in the sociological literature concerns differences in indicators of health status among men and women, in particular mortality differences. Some suggest that men's life styles and work activities (for example, hazardous work, social approval for smoking and drinking, and disapproval for accommodating illness) contribute to poor health; others emphasize that genetics and/or hormonal differences are the primary determinants of the differences. This debate focuses attention on the future of female health as the activity pattern of women moves toward that of men.

## II. Modelling the Determinants of Changing Health Status

Changes in health status over time are the result of a wide variety of exogenous characteristics and endogenous behaviors and statuses. By addressing change—a dynamic relationship—rather than level, we are able to control for the influence of many exogenous factors such as family and community background. We thereby focus on a set of variables involving the allocation of time, the choice of activities within any allocation, and other changes over the period, while controlling for the access to resources. This model allows both exogenous and endogenous factors to determine changes in health status.

Figure 1 is a simple description of the model which guides our empirical estimation. For empirical purposes, the "exogenous factors" are grouped into demographic, economic, and taste factors. Because our focus is on market work and its relationship to other time and activities allocations, we model hours of work to be dependent on the wage rate and on a variety of other exogenous factors. Both of these endogenous variables are imputed to represent the wage rate and hours of work if the individual had no health problem, that is, we remove health heterogeneity. The time and activity choice factors are similarly adjusted to represent hours allocated if the women had no early

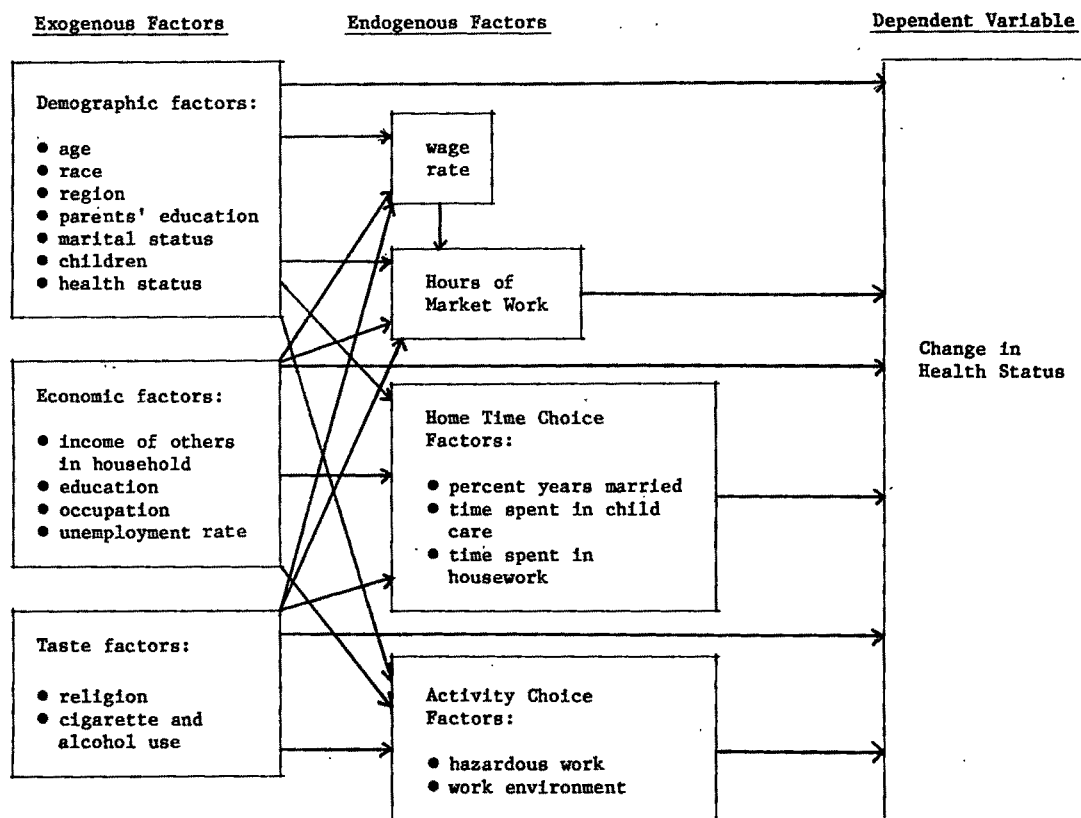


FIGURE 1

health problem. Finally, changes in health status are viewed as depending on selected demographic, economic, and taste factors, expected hours of market work if healthy, expected time allocations involving child care and housework if healthy, and choices of activity such as hazardous or stressful work. In our empirical work, we will emphasize the roles of time spent in market work, time spent in child care and housework (taken together as reflecting total nonmarket time demands) and exposure to a variety of hazardous jobs and environments in the market work place.

### III. Time Allocation, Hazards, and Changes in Female Health Status: Empirical Estimates

Empirical estimates consistent with the model of Figure 1 suggest that 1) market

work time, by itself, appears to contribute positively to changes in health status, 2) market work in hazardous jobs contributes negatively and significantly to health changes, 3) home time demands (defined as the sum of times spent in child care and housework) contributes negatively to changes in health, 4) dual role time demands (defined by percent of years worked with young children) contributes negatively and significantly to changes in health, 5) being self-employed contributes to health deterioration (perhaps because of stress), and 6) cigarette smoking contributes negatively and significantly to health status changes.

These estimates are presented in Table 1, as an *OLS* regression of the determinants of changes in female health (defined as the change in self-reported disability) from 1970 to 1976, and a trichotomous logit maximum-

TABLE 1—DETERMINANTS OF THE CHANGE IN WOMEN'S HEALTH,<sup>a</sup> 1970–76

Determinants <sup>b</sup>	OLS	Trichotomous Logic <sup>c</sup>		$\bar{X}$	$\sigma$
		Health Improve	Health Deteriorate		
Aggregate Market Work Hours <sup>d</sup>	-.21 <sup>-4</sup> (5.03)	.88 <sup>-4</sup> (2.24)	-.15 <sup>-3</sup> (4.92)	5689	5315
"Dual Role" Time Demand Factors					
Average annual child care plus housework hours <sup>e</sup>	.17 <sup>-4</sup> (1.42)	-.30 <sup>-3</sup> (2.15)	.16 <sup>-3</sup> (1.83)	1620	900
Percent of years worked with small children <sup>e</sup>	.09 (2.70)	-.67 (2.14)	.51 (2.16)	.19	.33
Percent of years married <sup>e</sup>	-.08 (2.82)	1.30 (5.44)	-.68 (4.03)	.74	.40
No market work 1970–1976	-.03 (1.15)	.12 (0.57)	-.13 (0.89)	.23	.42
Work Environment Factors					
Aggregate hours in jobs indexed by extent of physical hazards <sup>d</sup>	.25 <sup>-6</sup> (1.79)	-.27 <sup>-5</sup> (2.11)	.31 <sup>-5</sup> (2.91)	64366	99243
Aggregate hours in jobs indexed by extent of environmental hazards <sup>d</sup>	.23 <sup>-6</sup> (1.26)	-.30 <sup>-5</sup> (1.67)	.26 <sup>-5</sup> (1.87)	1716	57796
Stress Factors					
Change in marital status	.11 <sup>-2</sup> (0.04)	.47 (2.63)	-.07 (0.54)	.15	.36
Percent of years self-employed	.23 (1.66)	<i>t</i>	<i>t</i>	.01	.06
Taste Factors					
Cigarette use (number)	.84 <sup>-2</sup> (2.08)	-.31 <sup>-3</sup> (0.86)	.44 <sup>-3</sup> (1.72)	394	453
Constant	.019 (0.20)	-1.35 (1.55)	-.54 (0.92)		
$R^2$	.05				
$F$	6.12				

Note: *t*-statistics are shown in parentheses.

<sup>a</sup>Change in reported disability 1976 minus 1970:  $\bar{X} = .10$ ,  $\sigma = .42$ ,  $N = 2325$ .

<sup>b</sup>Other variables included in equation are family income minus own earnings (quadratic), work experience prior to 1970, education (quadratic), age, race, and region. The full specification of the models and the regressions on which the adjustments are based are available from the authors upon request.

<sup>c</sup>The vector of coefficients for the third option—no change in health status—is not reported.

<sup>d</sup>Adjusted to reflect variable value if health status in prior year equalled healthy.

<sup>e</sup>Adjusted to reflect variable value if health status in 1970 equalled healthy.

<sup>f</sup>Not included—no observations in improved category.

likelihood regression of the determinants of improvement, no change, and deterioration in health. The data are a sample of 2,325 women aged 25–65 in 1968, from the Michigan Panel Survey of Income Dynamics (PSID).

In order to control for the strong effect of early upon later health (and to purge time allocations of early health dependency), a three-stage estimation process has been employed in constructing the time demand variables. The market work time variable is an estimate of the aggregate number of hours

each woman would engage in market work over the seven-year period if she were healthy. This value is calculated by adjusting actual hours for the effect of health status on annual hours of work operating directly on hours worked and indirectly through the wage rate.<sup>1</sup> The effect is reflected in the coefficient

<sup>1</sup>The strength of the estimated negative effect of adjusted hours worked on the deterioration in health status suggests that our procedure may not have fully controlled for early health status. The crudeness of the health status variable (self-reported disability status) used for this adjustment may explain this result.

cients on average wage rate and disability status in annual hours worked regressions, 1970–76.<sup>2</sup> Obtaining hours worked if healthy involves adjusting actual hours by the product of the coefficients on disability status and average wage from these regressions, and assumed “healthy” levels of these variables. The average wage rate if healthy which enters this prediction is itself an adjusted value where the adjustment is based upon the coefficient on disability status in a regression of the *log* of the average wage rate (1970, 1976, in 1970 dollars).<sup>3</sup> Similarly, the variables indicating total nonmarket demands on time, years married, and the hours worked in hazardous jobs reflect adjusted time allocations to reflect the choices which would be made were each woman healthy.

A few of the other independent variables in Table 1 should also be mentioned. First, the hazardous work variables are the first two vectors of a principal components analysis based upon 10 job characteristic variables. These job characteristics were matched to 1970–76 *PSID* occupations, and are from information in the Dictionary of Occupational Titles. The first factor has high weights on job requirements such as stooping, climbing, strength and job conditions such as atmosphere and hazards. The second has high weights on environmental variables such

as noise, cold, wet, and heat. These account for 56.5 and 32.5 percent of the total variance, respectively.

Second, the past work history of each individual is captured by a variable indicating the total number of years of work experience of the individual at the beginning of the period of analysis (1970). In some respects, this variable also proxies for early health, but it is not used to adjust the (1970–76) time allocation variables. Third, the time demand variable includes hours spent in child care in 1976 (the only year reported) and an average over the seven-year period of hours spent in housework, the sum of these variables adjusted to represent time demands as though the woman was healthy at the beginning of the period. Together with average hours worked (if healthy) these represent total time demands. Fourth, beyond hours of time demands, working and having children less than 6 years old may create additional demands, and therefore a variable indicating the number of years over this period when a woman faces this dual role is included. Since self-employment may create time demands and stresses over and above those reported as work hours, the percentage of the 1970–76 period in which a woman is in this status is included. Finally, a variety of background and family status variables for the 1970–76 period are included.<sup>4</sup> They are primarily viewed as control variables to enable accurate measurement of the effect of time allocation on health status changes. Their coefficients, however, are interesting in their own right.

#### IV. Conclusions

Our results indicate that time allocations do have a significant effect on health status. In general, these results indicate that market work does not, in itself, cause health problems, and may in fact contribute to improvements in health. However, both the child care and housework demands on women and

<sup>2</sup>These regressions include: *log* of actual (if available) or predicted average wage rate, other family income, education, spouse education, 0–1 head of household, 0–1 whether married with children, unemployment rate, disability status (lagged 1 year), age of youngest child, number of children, age and age square, race, if spouse disabled, if parents were poor, occupation, region, religion, and a risk avoidance index.

<sup>3</sup>This regression includes, as independent variables: education, education square, 0–1 foreign born, number of children less than 18 years (1976), religion, number of children before age 25, age of the youngest child (1970), average number of children (1970–76), number of years of work experience (1970), experience (1970) square, occupation, disability status (1970), spouse disability status, mother's education, spouse education, race, percent of years married from 1970–76, region (1970), unemployment rate (1970–76), training time required for occupation, other family income, number of years married and with children (1970–76), and a Heckman selection correction variable (calculated from a probit regression explaining the presence of an observed average wage).

<sup>4</sup>These include region, age, age square, income other than that earned by the respondent, race, education, education square, an index of cigarette smoking, and change in marital status.

the dual role of working and having young children appears to be associated with health deterioration. And, apart from hours of market work or aggregate time demands, health deterioration is significantly related to the nature of the market work done and its environment and to activities (for example, cigarette smoking) often asserted to be detrimental to health.

These results, however, are only indicative. Several avenues for further work to improve confidence in the estimates are clear. They involve the development of more comprehensive measures of true health status, additional efforts to model the numerous endogenities in the process by which changes in health status are determined, extending the time period over which the analysis occurs, and including longitudinal information on numerous phenomena believed to be related to health status (access to health care resources, nutrition, and exercise) and information on the stress and psychological demands of market work and home circumstances (which demands are likely to influence both market and home time allocations and health status). These data are not likely to be available in the near future.

Nevertheless, the above results represent the only estimates available which seek to disentangle the time allocation-health status nexus for women, and their implications are important. These implications apply to national health care policies, private and social insurance actuarial adjustments for the quantity of female market work (ignoring its interaction with other time demands), and to intrafamily decisions regarding the aggregate

allocation of market and nonmarket responsibilities.

## REFERENCES

- Bartel, Anne and Taubman, Paul, "Health and Labor Market Success: The Role of Various Diseases," *Review of Economics and Statistics*, February 1979, 61, 1-8.
- Grossman, Michael, "On the Concept of Health Capital and the Demand for Health," *Journal of Political Economy*, March/April 1972, 80, 223-55.
- Hill, C. R. and Stafford, F. P., "Allocation of Time to Pre-School Children and Educational Opportunity," *Journal of Human Resources*, Summer 1974, 9, 323-41.
- \_\_\_\_\_ and \_\_\_\_\_, "Parental Care of Children: Time Diary Estimates of Quantity, Predictability and Variety," *Journal of Human Resources*, Winter 1980, 15, 219-39.
- Leibowitz, A., "Education and the Allocation of Women's Time," in F. T. Juster, ed., *Education, Income and Human Behavior*, New York: McGraw-Hill, 1975, 171-98.
- Verbrugge, L. M., "Recent Trends in Sex Mortality Differentials in the United States," *Women and Health*, Fall 1980, 5, 17-37.
- Waldron, I., "The Coronary-Prone Behavior Pattern, Blood Pressure, Employment and Socio-Economic Status in Women," *Journal of Psychosomatic Research*, 1978, 22, 79-82.
- \_\_\_\_\_, "Employment and Women's Health: An Analysis of Causal Relationships," *International Journal of Health Services*, 1980, 10, 435-54.

## LONG WAVES IN ECONOMIC ACTIVITY

### Long Waves and Technological Innovation

By EDWIN MANSFIELD\*

In recent years, there has been a renewed interest in long waves, a subject that received considerable attention from economists like Kondratiev and Schumpeter. The slowdown in economic growth during the 1970's and 1980's prompted this renewed interest, since some economists believe that, after the expansion of the 1950's and 1960's, we are now experiencing the recessionary phase of the Kondratiev cycle. The existence of long waves is, of course, controversial. Some, like Christopher Freeman and Jay Forrester, believe in their existence; others, like Paul Samuelson, regard them as "science fiction."

At least since the days of Schumpeter, long waves have been associated with innovation. Although the past twenty-five years have seen a very significant increase in the amount of empirical research carried out by economists on various aspects of technological change, only a small amount of this research has been devoted to the timing of innovations. My assignment in this paper is to describe and review (and in a few cases, try to extend) what little we know about three interrelated questions: (1) Are there long waves in innovation? (2) Has the innovation rate declined in the 1970s and 1980s? (3) Do depressions tend to trigger innovations?

#### I. Are There Long Waves in Innovation?

Because of the enormous difficulties in defining, dating, and weighting inventions and innovations, there is very little information concerning the extent to which they cluster together and whether these clusters (if they exist) occur about forty to sixty years

apart. The best known work is that of Gerhard Mensch (1979) who has assembled data concerning the number of "basic" innovations introduced during each decade from 1740 to 1960. Mensch regards basic innovations as those that "are the source from which new products and services spring and in turn create new markets and new industrial branches to supply them" (p. 122). Mensch's data seem to indicate some clustering around 1770, 1825, 1885, and 1935. According to George Ray (1980), each cluster of innovations seemed to occur about ten to twenty years before the trough of a Kondratiev long wave.

However, as Mensch probably would be the first to point out, these data should be viewed with the utmost caution. What he is trying to do is to date the beginning of new industries, which is a very difficult job. To illustrate how hard it sometimes is to date a basic innovation, consider the case of the diesel locomotive. Whereas Mensch dates it at 1934, one could argue that an equally plausible date is 1913, two decades before the General Motors locomotive appeared. In addition, it is not clear how basic innovations should be (or are) distinguished from other innovations. For example, Mensch does not include the electronic computer or the birth control pill as basic innovations, but does include the zipper. Moreover, the reasons for excluding the many important, but not basic, innovations are not obvious, at least to me. Further, one wonders whether some attempt should not be made to weight basic innovations since they undoubtedly were of quite unequal importance.

For these and other reasons, there has been considerable criticism of these data. John Clark, Christopher Freeman, and Luc Soete (1981) emphasize that Mensch's data rely heavily on the sample of inventions in

\*University of Pennsylvania. My research was supported by a grant from the National Science Foundation.

TABLE 1—PATENTS GRANTED IN THE UNITED STATES BY NATIONALITY OF INVENTOR, 1966–77

Year	Total	Granted to Nationals	Granted to All Foreigners	Foreign Patents Granted to U.S.
1966	68,406	54,634	13,772	49,098
1967	65,652	51,274	14,378	47,982
1968	59,102	45,782	13,320	48,229
1969	67,557	50,395	17,162	50,852
1970	64,427	47,073	17,354	48,807
1971	78,136	55,988	22,328	49,849
1972	74,818	51,515	23,293	49,628
1973	74,139	51,501	22,638	43,326
1974	76,275	50,643	25,632	39,990
1975	71,994	46,603	23,391	39,300
1976	70,236	44,162	26,074	38,028
1977	65,269	41,383	23,886	39,477

Source: National Science Foundation (1981).

John Jewkes et al. (1958), which was not claimed to be a statistically reliable sample. They point out that, because the book was done in the mid-1950's, it could not have done justice to the innovations in the 1950's or 1960's. Also, since Jewkes et al. were concerned with inventions, not innovations, there may well be a tendency to underestimate the number of innovations at the beginning of this century (see Peter Senge, 1982). Further, it is not clear why Mensch excludes about a dozen of the Jewkes et al. cases (such as bakelite and stainless steel), apparently because they had a relatively short lag between invention and innovation.

Given the many criticisms of these data, I think that it is fair to conclude that, although they are interesting, they must be supplemented with a great deal of additional work before any reasonably firm conclusions can be drawn from them. Until such work is carried out, it is difficult to take other than a rather agnostic stance concerning the existence of Mensch's long waves.

## II. Has there been a Decline in the Rate of Innovation?

Some proponents of long waves in economic activity believe that the rate of technological innovation has been falling in recent decades. To what extent is such a belief supported by the available evidence? Without question, the sharp reduction in the rate of

productivity growth in the United States (and many other countries) during recent years is consistent with such a view. But a variety of factors (including declines in the rate of increase of the capital-labor ratio, the huge increases in oil prices, and the impact of various kinds of regulations) could have been partly responsible. Certainly the fact that there has been a productivity slowdown does not prove by itself that there has been a slowdown in the rate of innovation.

The patent rate in the United States fell after 1971 (Table 1). In practically all of the 52 product fields for which data are available, the number of patents granted annually (by year of application) to U.S. inventors declined during the 1970's. Further, a decline in the patent rate seems to have occurred too in many other countries, such as Germany, France, the United Kingdom, and Canada (but not Japan). However, the crudeness of patent statistics as a measure of the rate of innovation should be emphasized. The average importance of the patents granted at one time and place may differ from those granted at another time and place, and the proportion of total inventions that are patented may vary significantly. Also, patents are generally used to measure the rate of invention, not the rate of innovation. (See my article with M. Schwartz and S. Wagner, 1981.)

In the industries where one can measure the number of major innovations that are carried out per unit of time, there seems to

be direct evidence of a fall in the rate of innovation. For example, in the pharmaceutical industry, the number of new chemical entities introduced per year in the United States has declined relative to the 1950's and early 1960's. In the pesticide industry, too, there was a decline during the 1970's in the number of new products marketed per year. These measures suffer from the fact that it is difficult to find suitable weights for different innovations. Also, these measures overlook the small innovations that sometimes have a bigger cumulative effect than some of the more spectacular innovations. But nonetheless the results are of interest.

Thus, many of the available bits and scraps of data point to a slackening of the rate of innovation in the United States. But the data are so crude and incomplete that it would be foolish to put much weight on them. Also, it is important to distinguish between various sectors of the economy. In pharmaceuticals and agricultural chemicals, there may very well have been a decrease in the rate of innovation, due partly to increases in regulatory requirements. But in other parts of the economy, such as electronics and telecommunications, the rate of innovation seems to be very impressive. For example, recent advances in microprocessors and microcomputers are of great importance. Without denying that a slackening of the rate of innovation may have occurred in some industries, it seems to me that there is little evidence of such a decrease in other important industries. Overall, there may have been a slackening, but I know of no serious attempt in the United States to measure its size with reasonable precision.<sup>1</sup>

### III. Do Major Depressions Trigger Innovations?

Many proponents of long waves seem to believe that the innovation rate was relatively high during the Great Depression of the 1930's. Some, such as Mensch, believe

that major depressions tend to trigger and accelerate innovations. This is not a new idea. For example, William Brown (1957) and Ruth Mack (1941) have argued that new designs tend to be postponed during good times and that ideas that accumulate are tried out, and new ones explored, during depressions. In recent years, Mensch's statement and defense of this thesis have provoked considerable controversy.

In particular, Clark, Freeman, and Soete, while acknowledging that there may have been a clustering of major innovations during the 1930's, deny that it was due to the depression. Instead, they attribute it to quite different causes. In the case of synthetic materials, they point to the advances in basic science due to the work of Staudinger on the structure of long chain molecules and the pressure on the demand side due to German rearmament (especially in connection with synthetic rubbers). Further, they point out that, because of technical links, one innovation often leads to others.

With respect to some industries, like iron and steel, petroleum refining, and bituminous coal, what little data we have do not indicate that the innovation rate was relatively high when capacity utilization rates were very low. Based on data regarding 175 innovations in these industries during 1919-58, I estimated the relationship in each industry between the number of innovations in a particular year and the industry's capacity utilization rate in that year. Up to some point (about 70 percent of capacity in each of these industries), increases in the rate of utilization of capacity seemed to be associated with increases in the rate of occurrence of process innovations; beyond that point they were associated with decreases in their rate of occurrence.<sup>2</sup>

<sup>1</sup>It is important to point out that there were a number of reasons why, based on existing econometric models of R&D, innovation, and productivity, one would have expected some decrease in the rate of innovation in the 1970s and 1980s. See Mansfield et al. (1982).

<sup>2</sup>These results were based on a quadratic equation. All of the regression coefficients were significant at the 0.10 level, and most were significant at the .05 level. For a description of the data, see my 1968 study. Needless to say, this analysis and these data are very rough. For one thing, as pointed out earlier, it sometimes is difficult to date innovations, although the problems of this sort for these innovations seem much less severe than for some of the basic innovations discussed there.

According to executives in these industries, new processes are unlikely to be introduced when an industry is operating at low levels of capacity utilization, because the risks involved seem particularly great under those circumstances, profits being slim and the future seeming particularly uncertain. On the other hand, when an industry is operating at very high levels of capacity utilization, there is some reluctance to innovate because it will interfere with production schedules: there is little unutilized capacity that can easily and cheaply be used for "experimental" purposes.

Turning to product innovations, there was no statistically significant evidence that the rate of innovation in these industries varied appreciably over the business cycle. However, this may have been due in part to "noise" in the basic data, which are probably not as accurate as those for process innovations. Also, it should be recognized that most of both these process and product innovations are of less importance than those studied by Mensch, who focused on innovations that resulted in the formation of new industries and the revolutionizing of old ones.

Although these and other available data are far from adequate, I share the skepticism of Clark, Freeman, and Soete concerning the proposition that severe depressions trigger and accelerate innovations. One could argue that neither severe depression nor rapid inflation is conducive to major innovations. When sales are depressed and the future looks grim, the climate for innovation is hardly bright. And when double digit inflation occurs, this reduces the efficiency of the price system as a mechanism for coordinating economic activity and discourages investments in innovation and long-term R&D (see my 1980 article). But although this view is consistent with the above results concerning process innovations in the steel, oil, and coal industries, the limitations and tentativeness of these results should be stressed.

#### IV. Conclusion

My review of the published evidence does not persuade me that the the number of major technological innovations conforms to

long waves of the sort indicated by Mensch's data. Although it is likely that some clustering occurs in the number of innovations, the evidence supporting the view that well-defined clusters occur every forty to sixty years seems limited and open to serious criticism. Much more work of both empirical and theoretical sorts is needed if we are to understand whether there are long waves in innovation or invention, and if so, why they occur. At present, very little work of this sort is going on (particularly in the United States). In part, this is due to the very severe measurement problems that are encountered in this specific area, as well as questions concerning the existence and explanations of long waves in economic activity in general.

The hypothesis that severe depressions trigger and accelerate innovations is also questionable. To obtain further information on this score, more attention should be devoted to the timing of innovations and the effects of macroeconomic conditions and policies on innovation and technological change (as well as the reverse effects). Many models totally ignore the links between the macroeconomic climate, on the one hand, and the rate of innovation and productivity change, on the other. Despite all of the evidence to the contrary amassed in the past twenty-five years, there is still a tendency in some quarters to view innovation and technological change as exogenous to the economic system (or linked to it in an oversimplified fashion). This is very unfortunate, both from the point of view of economic analysis and policy formulation.

#### REFERENCES

- Brown, William H., "Innovation in the Machine Tool Industry," *Quarterly Journal of Economics*, August 1957, 17, 406-25.
- Clark, John, Freeman, Christopher and Soete, Luc, "Long Waves, Inventions, and Innovations," *Futures*, August 1981, 4, 308-22.
- Jewkes, John, Sawers, David and Stillerman, Richard, *The Sources of Invention*, London: Macmillan, 1958.
- Mack, Ruth, *The Flow of Business: Funds and Consumer Purchasing Power*, New York: Columbia, 1941.

- Mansfield, Edwin, *Industrial Research and Technological Innovation*, New York: W. W. Norton, 1968.
- , "Research and Development, Productivity, and Inflation," *Science*, September 5, 1980, 209, 1091–93.
- , Schwartz, M. and Wagner, S., "Imitation Costs and Patents: An Empirical Study," *Economic Journal*, December 1981, 91, 907–18.
- et al., *Technology Transfer, Productivity, and Economic Policy*, New York: W. W. Norton, 1982.
- Mensch, Gerhard, *Stalemate in Technology*, Cambridge: Ballinger, 1979.
- Ray, George, "Innovation in the Long Cycle," *Lloyds Bank Review*, January 1980, 14–28.
- Senge, Peter, "The Economic Long Wave: A Survey of Evidence," unpublished paper, Massachusetts Institute of Technology, April 1982.
- National Science Foundation, *Science Indicators 1980*, Washington: Government Printing Office, 1981.

# Long Waves and Economic Growth: A Critical Appraisal

By NATHAN ROSENBERG AND CLAUDIO R. FRISCHTAK\*

No one who has examined the dynamics of capitalist economies over long historical periods can doubt that they experience significant long-term variations in their aggregate performance. The question is whether these long-term variations are more than the outcome of a summation of random events, and further, whether they exhibit recurrent temporal regularities that are sufficiently well-behaved to call them "long waves." In recent years there has been a strong resurgence of interest in such long-term movements, since their existence could provide a coherent explanation for the poor performance of capitalist economies over the past decade.

We do not attempt herein to examine the historical evidence for long cycles. We have in fact examined this evidence and find it unconvincing. Although historical data might conceivably lend some plausibility to the notion of long cycles in prices, we remain, at present, sceptical of the case that has so far been made for their presence in real phenomena; that is, in aggregate output or employment.

What we offer here, instead, consistent with our present scepticism, is an attempt to examine the economic logic of long waves. More specifically, we ask what conditions would need to be fulfilled in order for technological innovation to generate long cycles of the periodicity postulated by N. D. Kondratiev and his disciples? It is our view that such a theory, which might account for the presence of long cycles in some real economic variable, would have to fulfill a set of logically interdependent requirements. We discuss these requirements under the four categories of causality, timing, economywide repercussions, and recurrence.

## I. Causality

The first of these requirements is a clear specification of causality among the factors associated with long wave phenomena. Kondratiev was insistent that capitalism had its own internal regulating mechanisms, and he regarded the pace or rhythm of the long cycle as an expression of these internal forces. Long cycles, as Kondratiev put it, "...arise out of causes which are inherent in the essence of the capitalist economy" (1979, p. 543). The cyclical behavior of the capitalist economy in turn shapes the conditions that are favorable to technological innovation. In this specific sense, therefore, technological activities stand in the position of dependent variables whose volume and timing are determined by those deeper-rooted forces that shape the rhythm of capitalist development. In spite of substantial differences among them, several present-day advocates of long cycles—W. W. Rostow, E. Mandel, and Jay Forrester—seem to share the Kondratiev view that innovations are, somehow, disciplined and structured, and have their timing determined by such long-term movements.

Schumpeter was, of course, the foremost and most influential articulator of the opposite view—that long cycles are caused by, and are an incident of, the innovation process. In Schumpeter's view, innovation is at the center of both cyclical instability and economic growth, with the direction of causality moving clearly from fluctuations in innovation to fluctuations in investment and from that to cycles in economic growth. Moreover, Schumpeter sees innovations as clustering around certain points in time—periods that he referred to as "neighborhoods of equilibrium," when entrepreneurial perception of risk and returns warranted innovative commitments. These clusterings, in turn, lead to long cycles by generating periods of acceleration (and eventual deceleration) in aggregate growth rates. Why cluster-

\*Stanford University. We thank Moses Abramovitz for his careful criticisms of an earlier version.

ing should occur is obviously crucial to a theory of long cycles, and we will therefore return to this question shortly. But it is essential to stress that an exponent of the view that technological change is at the root of the long cycle needs to demonstrate (a) that changes in the rate of innovation govern changes in the rate of new investment and (b) that the combined impact of innovation clusters takes the form of fluctuations in aggregate output or employment.

## II. Timing

The process of technological innovation involves extremely complex relations among a set of key variables—inventions, innovations, diffusion paths, and investment activity. A technological theory of long cycles needs to demonstrate that these variables interact in a manner that is compatible with the peculiar timing requirements of such cycles.

It is not enough to argue that the introduction of new technologies generates cyclical instability. It is necessary to demonstrate why technological innovation leads to cycles of four and a half to six decades in length, with long periods of expansion giving way to similarly extended periods of stagnation. We confine ourselves here to a brief inventory of strategic factors that may be expected to determine the length of the time period required for the introduction of new technologies and the realization of their full impact upon aggregate output.

New inventions are typically very primitive at the time of their birth. Their performance is usually poor, compared to existing (alternative) technologies as well as to their future performance. Moreover, the cost of production at this initial stage is likely to be very high—indeed, in some cases a production technology may simply not yet exist, as is often observed in major chemical inventions (nylon, rayon). Thus, the speed with which inventions are transformed into innovations, and consequently diffused, will depend upon the actual *and* expected trajectory of performance improvement and cost reduction.

This process is rendered more complex, first, by the fact that in the early stages, when performance is still very modest and production costs are high, improvements leading even to significant cost reductions may have no sizable effect upon rates of adoption, often resulting in long gestation periods. When, on the other hand, the new product attains cost levels roughly equivalent to those prevailing under the older technology, even *small* further cost reductions may lead to widespread adoption. Thus, there may be a highly nonlinear relationship between rates of improvement in a new product and rates of adoption.

Second, since innovation and investment decisions are future oriented and therefore inevitably involve a high degree of uncertainty, adoption and diffusion rates are also powerfully shaped by expectation patterns. In certain cases, these expectation patterns may lead to a prolonged delay in the introduction of potentially superior new technologies by adversely affecting their expected profitability. Indeed, it has been very common for competitive pressures generated by a new technology to lead to substantial improvements in the old technology, so that the new one established its superiority more slowly than would otherwise have been the case.

Note in this respect that major improvements in productivity may be stretched out over very long time periods, as a product goes through innumerable minor modifications and alterations in design. (Consider the Fourdrinier papermaking machine, which has now dominated the paper industry for almost 200 years.) To the extent that major innovations vary with respect to the relevant time period for which they remain important, in part because substantial improvements will often take place long after the initial introduction of the innovation, it renders highly problematical the whole exercise of inferring a Kondratiev long cycle from a particular innovation. How does one date the long cycle associated with the steam engine? Beginning with Watt's seminal inventions in the 1770's? What we know about the slow pace of its adoption in the late

eighteenth century renders this extremely doubtful. But, in addition, the improvements associated with the compound engine brought huge productivity improvements sufficient to introduce the steam engine to important new uses, and this came a full century after Watt's major contribution.

Third, the adoption of a new technology is often critically dependent upon the availability of complementary inputs or, in some cases, upon an entire supporting infrastructure. Automobiles required extensive networks of roads, gasoline stations, and repair facilities. The electric lamp required an extensive system for the generation and distribution of electric power. Seldom do new products fit into the existing social system without some intervening period of accommodation during which these complementary considerations are arranged. Not only does this signify a heavy commitment to an established technology, and a further reason for a *slow* initial shift to a best practice frontier, but, moreover, the time period required for such accommodations may vary greatly from one innovation to another.

Even if major innovations experience appropriately long and logistically shaped diffusion paths, with technologies going through phases of accelerated growth and eventually petering out, it does not necessarily imply that the slack in the declining phase of an individual innovation cycle might not be taken up by other technological innovations, thus eliminating the impact of a long phase of "sectoral" retardation. What would still be needed for a wave-like pattern of growth is that other major substitute innovations were excluded until the original one had run its course. Without such a *spacing* mechanism, partially overlapping innovations might otherwise generate steady rates of growth rather than cycles.

What technological forces might impose cyclical behavior rather than some sort of relative stability of economic activity along any given path traced out by a sequence of major substitute innovations? We have already suggested three such forces that might delay the introduction and widespread adoption of a new substitute technology, namely,

a production cost differential that may still persist between the old and new technologies; certain expectational patterns held in common by entrepreneurs regarding improvements in both technologies; and the cost associated with scrapping and replacing the infrastructure committed to the old one. An additional possibility is that major innovations may establish certain trajectories of readily available performance improvements and cost reduction (more circuits on a chip, fewer pounds of coal per kilowatt hour of electricity). Engineers and technically trained personnel often work with such implicit notions. Thus, the awareness of these trajectories may serve as focusing devices that fix the attention of engineers upon teasing out the further improvements that are understood to be available from the existing technological framework, rather than searching for entirely new technologies. In addition, these trajectories may be expected to shape the educational system and the training of engineers and other technical personnel. The inertial forces here may strengthen the commitment to an existing technology and render more difficult the exploration of new realms of technical possibilities.

The reasons so far invoked for lengthy delays in the adoption of new technologies, of a kind that might produce extended periods of industrywide stagnation, were discussed in connection with major *substitute* innovations. Should similar considerations of spacing be extended to cases of *unrelated* or *complementary* technologies unfolding along many different trajectories?

In the case of unrelated technologies, the answer, *prima facie*, would be no. Even if one argued that there were forces leading to the spacing of innovations in the same industrial sector, in the sense that the arrival of a new technology has to wait until the benefits of moving along the previous technological trajectory had been largely exhausted, this would be of limited relevance for major innovations in *other* industries. The fact that we are still on a highly productive portion of the steam trajectory might conceivably tell us something about the timing of substitute innovations, such as electric motors, but little

if anything about the timing of unrelated innovations and their subsequent diffusion in electronics, synthetic fibers, or pharmaceuticals.

Yet the long cycle hypothesis might be considerably strengthened if a large number of unrelated innovations had the main phases of their life cycles synchronized by macroeconomic conditions. Indeed, the simultaneous diffusion of a large number of unrelated innovations may be understood as being regulated by general conditions in financial, factor and product markets. If favorable, they might lead to a "bandwagon effect" along a number of separate industry trajectories, followed eventually by a slowdown. The result would be an innovation cluster of type *M*, the vertical summation of sectoral logistics.

In the case of related technologies, an additional reason can be invoked for the synchronization of separate diffusion paths: they may be linked by a system of technologically connected 'families' of innovations, made up of complementary, induced and closely related ones.<sup>1</sup> This would come about because the interactions of a few basic technologies would provide the essential foundations for other technological changes in a series of ever widening concentric circles. A technological cluster, or a cluster of type *T*, arises therefore when one (or a small number of) major related innovations provide the basis around which a large number of further cumulative improvements are positioned.

Yet, the question persists: is spacing within and synchronization among different diffusion paths, for both technological and macroeconomic reasons, sufficient to provide a long-wave pattern of aggregative growth? And if so, how?

It is our present, tentative assessment that modes of argument at the technological level, while potentially interesting and well worth further exploration, will be of only limited usefulness in providing a convincing account of the generation of long waves. Technologically driven long waves can be made to appear plausible only if macroeconomic factors can be shown to play a dominant role in

shaping and disciplining the timing of the introduction of innovations. The beginning of an upswing would therefore be characterized by a sufficiently large stimulus from the *M* clustering process upon the previously positioned *T* cluster. On the other hand, once activated by the macroeconomic environment, the technological long cycle is required to detach itself from swings in demand which closely track short-run changes in macroeconomic conditions, and instead follow the internal dynamics of technological factors. As we have already shown, these may lead to extended periods of multisectoral growth and retardation, although there is no reason to believe that they will add up to cycles of forty-five to sixty years' duration. What has not been shown so far is the connection between such factors and the derived and induced demands for capital and consumer goods which would account for the economywide impact of innovation clusters.

Therefore, to argue effectively for a technologically driven long cycle, an additional requirement has to be met: the cluster of innovations must occupy a strategic position in the economy in terms of backward and forward links.

### III. Economywide Repercussions

An essential step in a technological theory of long cycles is the demonstration of the mechanisms through which particular changes in technology exercise *sizable* changes in the performance of the macroeconomy. The economywide impact of technological innovations needs to be understood not only in terms of the direct impact of cost reductions and the release of resources to alternative uses, but of the strength of their backward and forward linkages.

1) They should be strongly linked backwards in terms of expenditures for buildings, machinery, equipment, and raw materials, such that the initial innovation and investment requirements lead to further investment decisions in the production goods sector. Historically, this second wave of investment has often bred a second wave of innovations, more explicitly "process" oriented, and concentrated in the production goods sector. It

<sup>1</sup> On this last point, see C. Freeman et al., 1982, ch. 4.

should be particularly noted that this last set of innovations has frequently had the effect of increasing the productivity of the economy at locations far from the specific sector that originally gave rise to the innovative activity.

2) The impact of innovations will also depend upon the strength of their forward linkages. These might take the form of a reduction in the price of the products into which the innovation enters as an input, leading to an expansion in the size of their market, and therefore also to an expansion in the rate of capital accumulation, output growth, and technical progress in these industries. These induced responses would depend upon the number of industries into which the innovation enters as an input, its substitutability for other inputs, the proportion of total costs it accounts for and the extent of cost reductions it imposes upon the product.<sup>2</sup> More important, innovations may induce the creation and diffusion of new products and processes that, in their turn, would bring about the widespread adoption of the original innovation (the microchip is a compelling recent example). Alternatively, the impact will depend upon the extent to which the initial innovation proves to be at the core of "major natural trajectories" (such as the electric motor in relation to the process of electrification), or more generally, in key sectors of the economy, such as energy and transportation.

In sum, the interindustry flow of new materials, components and equipment may generate widespread product improvement and cost reduction throughout the economy. This has clearly been the case in the past among a small group of producer goods industries—machine tools, chemicals, electrical and electronic equipment. Industrial purchasers of such producer goods experienced considerable product and process improvement without necessarily undertaking any research expenditure of their own. Such interindustry flow of technology is one of the most distinctive characteristics of advanced

capitalist societies, where innovations flowing from a few industries may be responsible for generating a vastly disproportionate amount of technological change, productivity improvement and output growth in the economy. It is certainly conceivable that innovations, depending upon their location, generate long cycles through such interindustry flows and their consequent macroeconomic effects. Yet, given the difficulties of knowing what is the nature of the benefits flowing from each innovation, and where exactly within the structure of the economy these benefits eventually accrue, this can at best be regarded as no more than an untested hypothesis until systematic attempts at quantification have been undertaken.

#### IV. Recurrence

The final requirement for a theory of long waves based upon technological innovations involves demonstrating their cyclical or recurrent character. In fact, it is not sufficient to show that causality runs from innovation to investment; that the economic and technological factors which determine the adoption of new technologies do so in a manner compatible with the stringent timing requirements of a Kondratiev; and that patterns of diffusion and interindustry linkage of new technologies involve sufficient amplitude that long cycles are perceived in the form of sizable variance in aggregative growth rates. It still has to be shown, if the argument is going to be logically complete, that the waves repeat themselves over time, either because the wave-generating factors in the form of innovation clusters are themselves cyclical (or at least recur with a certain regularity), or because there is an endogenous mechanism in the economic system which necessarily and regularly brings a succession of turning points.

Let us briefly examine the conditions under which long cycles become a historical necessity, in the sense that there are structural reasons for one long wave to follow another. The following should be present: 1) the availability of an elastic supply of inventions, at a time when risk-return combinations appear propitious for innovations; 2)

<sup>2</sup>See, in this respect, Albert Fishlow's book, which is a rigorous and imaginative attempt to measure the economywide impact of a single innovation.

the formation of a cluster of innovations at the base of the upswing, that is, a technologically dense set which undergoes a rapid process of diffusion under favorable macroeconomic conditions; 3) the reaching of an upper turning point of the technologically driven cycle due to increasing macroeconomic instability, as well as forces that deter the introduction of substitute technologies; 4) the arrival of the economy at a technologically fertile ground, after an appropriately extended period of time. At this point, old innovation paths have been largely exhausted, but previously postponed ones have not yet been taken up.

This schema brings numerous problems. One would be hard pressed to show that Kondratievs are regulated on a purely internal basis, and that, in the past, exogenous factors have had only a marginal effect upon such long-term movements. Others have in fact argued that the recurrence of innovation clusters has been more in the nature of *historical accidents* than endogenously-generated fluctuations in the rate of innovation. Further, our earlier discussion of timing provides no compelling reason to expect recurrence at forty-five to sixty-year intervals, even if innovations cluster and such clusters appear regularly.

What we have attempted to show is the far-from-trivial requirements that are necessary in order to demonstrate that technological change, in conjunction with macroeconomic forces, can indeed be the preponderant force behind long waves. Hav-

ing made these conditions explicit, we feel we are entitled to conclude that the conceptual framework of a model of long waves in economic growth, which has at its core the process of technological innovation, has still not been adequately formulated. Until such a model is developed, the assessment of its historical validity remains unresolved.

## REFERENCES

- Fishlow, Albert, *American Railroads and the Transformation of the Ante-Bellum Economy*, Cambridge: Harvard University Press, 1965.
- Forrester, Jay, "Innovation and Economic Change," *Futures*, August 1981, 13, 323-31.
- Freeman, C. et al, *Unemployment and Technical Innovation: A Study of Long Waves and Economic Development*, London: Frances Pinter, 1982.
- Kondratiev, N. D., "The Long Waves in Economic Life," *Review*, Spring 1979, 4, 519-62; a complete translation of "The Major Economic Cycles," *Voprosy Kon' iunktury*, 1925, 1, 28-79.
- Mandel, E., *Late Capitalism*, London: New Left Books, 1975.
- Rostow, W. W., *The World Economy: History and Prospect*, Austin: University of Texas Press, 1978.
- Schumpeter, J., *Business Cycles, A Theoretical, Historical and Statistical Analysis of the Capitalist Process*, New York: McGraw-Hill, 1939.

# Long Swings and the Nonreproductive Cycle

By DAVID M. GORDON, THOMAS E. WEISSKOPF, AND SAMUEL BOWLES\*

The U.S. and world capitalist economies are currently in the midst of the third long swing crisis of the past century. A notable and, we shall argue, defining characteristic of this and prior long swing crises is the failure of the business cycle, through its normal functioning, to restore conditions for rapid accumulation. We develop a theoretical model of the relation of the business cycle to long swing expansions and crises, and present some evidence supporting our hypotheses about the relationship between the long swing crisis and what we term the nonreproductive cycle. Our approach implies that the most theoretically coherent basis for dating long swings in capitalist economies is not to be found in the movements of total output or investment, but rather in an investigation of what Gordon has called the social structure of accumulation and its ability to restore profitability during cyclical downturns.

This model of the relationship of cycles to swings not only helps clarify the characteristic patterns of long swings, but may also help resolve a number of anomalies which have occasionally puzzled economists. The considerable rise in real wages during the sharp cyclical downturns of the 1890's and 1930's is no longer anomalous, for example, but is rather an expected feature of business cycles in a period of long swing crisis. (This does not explain, of course, why real wages rose.) Further, the somewhat inconclusive debate in the Keynesian literature about the procyclical or anticyclical behavior of real wages may potentially be clarified by our observation that real wages tend to move procyclically during long swing expansions and anticyclically during long swing crises—and that for this reason the prevailing debate has been

clouded by an underspecification of the determinants of cyclical behavior during long swings.

These anomalies are not surprising. Most studies of the relationship of the business cycle to long swings in economic activity have failed largely because of the absence of any theory of the relationship between these two forms of economic fluctuation; the same models of investment and output determination are typically applied to both. Recent Marxian analyses of the dynamics of long swings offer the possibility of overcoming this weakness. (See Gordon, Richard Edwards, and Michael Reich, ch. 2, for a review.) These analyses have stressed the crucial importance of a periodically reconstituted set of institutions, called the social structure of accumulation (SSA), which provides the economic stability and moderation of political economic conflict essential for favorable profit expectations and therefore for rapid capital accumulation. These institutions include, for example, systems of labor management, the international monetary system, and structures mediating raw materials supply. Their erosion sets the stage for economic crisis.

Despite the promise of this approach, however, the "normal" operations of the business cycle have not been integrated into the broader analysis of the SSA and long swings. We begin by distinguishing the reproductive (or well-behaved) cycle from the nonreproductive (or perverse) cycle. The reproductive cycle is one in which a downturn in economic activity is corrected by the functioning of the cycle itself. We call this cycle reproductive because it endogenously restores conditions for rapid accumulation without requiring fundamental changes in the structure of the accumulation process. The nonreproductive cycle, by contrast, is one in which a downturn does not correct itself endogenously, and which therefore requires basic changes in the institutions that

\*New School for Social Research, University of Michigan, and University of Massachusetts-Amherst, respectively. The order of our names has been selected randomly. We thank Peter Alexander for research assistance, and Carol Heim and Robert Zevin for helpful comments and suggestions.

regulate the accumulation process and establish the conditions for profitability.

This distinction allows us to define the difference between long swing expansions and crises in a particularly simple manner. Long swing expansions are characterized by reproductive cycles, sustaining the effectiveness of the *SSA* in promoting profitability, investment, and growth. Long swing crises are characterized by nonreproductive cycles, leading to prolonged periods of economic stagnation or disaccumulation and eventually, if capitalism is to continue, to the construction of a new *SSA* capable of rekindling profitability, investment, and growth. The theoretical distinction between long swing expansions and crises thus resides in the reproductive or nonreproductive nature of the business cycle; slower economic growth or reduced accumulation are therefore a probable consequence of a long swing crisis rather than its defining characteristic.

These definitions presuppose a specific set of interconnections among the *SSA*, the expected profit rate, the cycle, and investment activity. The expected profit rate depends on the effectiveness of those institutions which make up a given *SSA*. The level and pattern of investment depends upon the *SSA* and the expected profit rate. Cyclical downturns are induced by a decline in the expected rate of profit. Reproductive cyclical downturns restore the expected rate of profit, and thus investment activity, while nonreproductive cyclical downturns do not. (We provide some econometric evidence for these propositions in our 1983 book, and in Weisskopf's 1979 article. One effort at explaining the emergence of the nonreproductive cycle in the post-World War II period is found in Bowles and Herbert Gintis.)

Why would a reproductive cycle become nonreproductive? We may examine the functioning of the reproductive cycle through an analysis of the determinants of the profit rate. Abstracting from taxes, we may represent the profit rate of the individual firm,  $r$ , as the product of the share of profits in firm value-added  $s_r$ , the ratio of output to utilized capital stock  $y_u$ , and the ratio of utilized to owned capital stock  $k^*$ , or

$$(1) \quad r \equiv s_r y_u k^*,$$

where

$$(2) \quad r \equiv \Pi / K_0; s_r \equiv \Pi / Y; y_u \equiv Y / K_u; \\ k^* \equiv K_u / K_0;$$

and  $\Pi$  is firm profits,  $K_0$  is the value of the firm's owned capital stock,  $Y$  is firm value-added, and  $K_u$  is the portion of the owned capital stock which is currently utilized.

The expected profit rate  $r_e$  will by parallel reasoning depend on the "full capacity profit rate"  $s_r y_u$ , and expected capacity utilization  $k_e^*$ , or

$$(3) \quad r_e \equiv s_r y_u k_e^*.$$

The expression in (3) suggests that a cyclical downturn may restore expected profits in either of two principal ways, by raising the profit share,  $s_r$ , or by raising the ratio of output to utilized capital,  $y_u$ . (Downturns do not generally raise expected capacity utilization.) The cyclical downturn may raise  $y_u$  by eliminating high-cost firms and by inducing the nonuse of high-cost processes within surviving firms. The effect of the cycle on  $y_u$  is closely associated with the effect of the cycle on competitive pressures among business and, by implication, with the business failure rate.

The effect of a cyclical downturn on the profit share involves a somewhat more complicated set of connections. It may be investigated by representing the profit share as unity minus what we call real unit labor costs, or

$$(4) \quad s_r \equiv 1 - [(w/p)/qe],$$

where  $w$  is the nominal wage,  $p$  is the price of output,  $q$  is output per unit of labor effort, and  $e$  is labor effort input per hour.

A cyclical downturn may lower the product wage ( $w/p$ ), raise output per unit of effort ( $q$ ), or raise labor effort per hour ( $e$ ). The effects on  $w/p$  and  $e$  are derived from the increased power of capital over labor associated with an increase in the size of the reserve army of the unemployed and the consequent increase in the cost to the worker of losing his or her job. (These effects are estimated econometrically in our 1983 paper, and in Juliet B. Schor and Bowles.) The

effect of the downturn upon  $q$  may arise because of the firm elimination and competitive pressure effects outlined above, or because of increased power of capital in determining work rules and technical changes, or for other reasons.

The nonreproductive cycle is one in which these restorative effects of the downturn on the profit share and the ratio of output to utilized capital fail to operate. The failure of these effects signals an erosion of the effectiveness of the institutions of the SSA and therefore the onset of a long swing crisis. Under what conditions might they fail to operate?

The cyclical downturn might not raise the ratio of output to utilized capital stock if high-cost firms and high-cost operations were not eliminated, or if the downturn was characterized by an especially severe contraction in sectors with relatively high output/capital ratios, or if the reserve army effects failed to raise output per effort unit or effort per hour, or for other reasons.

The cyclical downturn might not raise the profit share if producer price cutting outweighed wage cutting, giving rise to an increase in the product wage, or if the contraction failed to have positive effects on output per unit of effort or effort per labor hour.

We do not have space to present a model of price determination over the cycle which incorporates both mark-up and limit-pricing behavior: our model suggests, in brief, that the negative effects of economic downturns on product prices are likely to be limited.

The effect of the contraction on the other components of real unit labor costs—the nominal wage rate, output per effort unit, and effort per labor hour—may be derived from a microeconomic analysis of the capital-labor conflict over wage setting and work intensity as modeled in Bowles. For our purposes here, it is sufficient to observe that the effects of the cyclical downturn on the above determinants of real unit labor costs will be more favorable to capital, the more the contraction increases the expected duration of unemployment and the greater is the difference between the worker's current after-tax wage and the worker's expected level of unemployment insurance and other income-replacing payments from the government.

In an accounting sense, the effect of a cyclical downturn on the expected profit rate will obviously depend on the sum of the above contradictory effects. More substantively, the overall effect will depend on the manner in which the social structure of accumulation regulates product markets, labor markets, the labor process, international exchanges, state expenditures, and so forth. Elaboration of an adequate model of these relations cannot be pursued in such a short essay. (Gordon, Edwards, and Reich provide this kind of analysis for the labor process and labor market.)

We confine ourselves, much more simply, to a brief empirical demonstration that our hypothesized relationships between the business cycle and long swings are at least partly confirmed by the historical data. Because data on the capital stock and its utilization are inadequate for cyclical analysis until relatively recent years, we cannot explore the effects of cyclical downturns on the level of output per utilized unit of capital stock  $y_u$ . But available data do permit investigation of the movements of real unit labor costs in the U.S. economy for the period from 1890 to 1982. (Our pre-1948 data refer to the manufacturing sector alone; data for 1948 and later years are for the private business sector. Sources and methods are fully described in a data appendix, available from the authors.)

We use these data to classify cycles as reproductive or nonreproductive. A nonreproductive cycle is defined as one in which the ratio of the product wage to output per hour, or real unit labor costs  $[(w/p)/q_e]$  rises rather than falls between the business cycle peak and the year following the trough. (We measure changes from peak to one year after the trough to allow time for the full restorative impact of the downturns.)

Our results in column (3) of Table 1 indicate alternating periods of nonreproductive and reproductive cycles, with the nonreproductive or crisis periods spanning the years 1890–1903, 1926–37, and 1969 to the present—the cycles numbered 1–4, 11–12, and 19–21, respectively. There is also a noticeable long swing pattern in the data in column (3), as one can see by graphing its cycle values against time. (We omit the graph for reasons of space.) This impression is sus-

TABLE 1—THE RESTORATION OF PROFITABILITY IN CYCLICAL DOWNTURNS, 1890–1981

Cycle	Cyclical Downturn		Percentage Average Annual Change			Change in Unemployment Rate
			Product Wages	Output per hour	Real Unit Labor Costs	
	Peak	Trough	(1)	(2)	(3)	(4)
1	1890	1891	0.78	0.00	0.78	1.4
2	1892	1894	3.84	−0.70	4.54	15.4
3	1895	1897	1.63	−0.67	2.30	0.8
4	1899	1901	2.03	1.45	0.58	−2.5
5	1903	1905	−0.93	2.55	−3.48	0.4
6	1907	1909	−3.24	2.67	−5.91	2.3
7	1910	1912	3.05	4.66	−1.61	−1.3
8	1913	1915	−1.69	6.28	−7.97	4.2
9	1919	1922	5.87	10.83	−4.96	5.3
10	1923	1925	2.28	6.30	−4.02	0.8
11	1926	1928	4.29	3.32	0.97	2.4
12	1929	1933	1.50	1.35	0.15	21.7
13	1937	1939	2.40	3.72	−1.32	2.9
14	1944	1947	−2.94	−1.34	−1.60	2.7
15	1948	1950	3.92	4.46	−0.54	1.5
16	1953	1955	0.96	2.73	−1.77	1.5
17	1957	1959	2.57	3.13	−0.56	1.2
18	1960	1962	3.14	3.48	−0.34	0.0
19	1969	1971	2.33	2.17	0.16	2.4
20	1973	1976	1.17	1.07	0.10	2.8
21	1979	1981	0.85	0.45	0.39	1.8

Notes: The NBER peak-trough-peak cycle of 1918–1919–1920 was ignored as a cycle, since the respective unemployment rates for these years were 1.4, 1.4, and 5.2 percent; we designate 1919 as the peak instead. Cycle peaks are those identified by the NBER; troughs are one year after the NBER trough, except that if this is another peak year, the NBER trough itself is used.

tained by evidence of autocorrelation when the data in column (3) are tested against the null hypothesis of constancy over time. The Durbin-Watson statistic for such a test is 0.90, significant at 5 percent and substantially lower than for comparable tests on the between-cycle movements of data on producers' prices, aggregate output, or aggregate investment.<sup>1</sup> Our model of the effects of cycles on expected profits, in short, reveals clearer evidence of long swing behavior than other macroeconomic indicators on which economists have previously concentrated.

Additional analysis supports our hypotheses about the differences between reproductive and nonreproductive cycles. In the

reproductive cycle, we would expect a systematic inverse relationship between the peak-to-trough changes in unemployment and real unit labor costs: the greater the reserve army jolt, the larger the reduction in real unit labor costs during the downturn. In the nonreproductive cycle, given our hypothesized erosion of the SSA, we would expect a breakdown of this effect; we should therefore find no evidence of any systematic statistical relationship between the movements in unemployment and real unit labor costs. We have combined the reproductive cycles (5–10 and 13–18) in one group, and the nonreproductive cycles (1–4, 11–12, and 19–21) in another. Regressing the data in column (3) on the data in column (4), we find a negative and statistically significant coefficient for the group of reproductive cycles and accept the hypothesis of statistical independence for the nonreproductive group.

<sup>1</sup>We performed these tests on data for the wholesale price index, gross private domestic product, and gross private domestic investment. The Durbin-Watson statistics for these tests were 1.28, 1.37, and 1.12, respectively.

TABLE 2—GROWTH OF OUTPUT AND INVESTMENT IN THE U.S. ECONOMY, 1860's–1979<sup>a</sup>

	Boom	Crisis	Boom	Crisis	Boom	Crisis	Averages	
	1860's to 1892	1892 to 1899	1899 to 1929	1929 to 1937	1937 to 1973	1973 to 1979	Boom	Crisis
Gross Domestic								
Nonfarm Product	6.6	2.9	3.7	−0.8	4.2	2.9	4.8	1.8
Industrial Production	5.1	3.4	4.5	0.4	4.9	2.7	4.8	2.2
Gross Domestic Private								
Fixed Nonres. Investment	8.0	1.3	2.4	−4.7	4.7	2.8	5.0	−0.2

<sup>a</sup>All entries are percentage average annual changes in constant dollars.

TABLE 3—AVERAGE EFFECTS OF CYCLICAL DOWNTURNS<sup>a</sup>

	Cycles	Real Unit Labor Costs (1)	Unemployment Rate (2)
Crises: <sup>b</sup>	1–3	2.54	5.9
	11–12	0.56	12.0
	19–21	0.22	2.3
	Average	1.18	6.1
Expansions:	4–10	−3.91	1.3
	13–18	−1.02	1.6
	Average	−2.58	1.4

<sup>a</sup>Columns (1) (shown in percent) and (2) represent the averages for the respective cycle groups of the data presented in columns (3) and (4) in Table 1.

<sup>b</sup>We include for this comparison the cycle identified as the onset of the crisis, or the first nonreproductive crisis, whereas we excluded this cycle in our presentation of the data in Table 2.

Covariance analysis further confirms the hypothesis of differences in both intercepts and slopes between the two groups.

We can next explore the relationship between the data in Table 1 and conventional hypotheses about long swings in economic activity. The first nonreproductive cycle in each sequence in column (3) (cycles 1, 11, and 19) appears to precede by one cycle those periods which are generally characterized as crises or depressions.<sup>2</sup> We present in Table 2 the results of dating economic crises proper, as opposed to their onsets, as commencing after one nonreproductive cycle. (We have dated the end of the 1890's crisis at the NBER peak of 1899,

discounting the nonreproductive nature of the 1899–1903 downturn, since the unemployment rate actually fell during this “contraction.”) Data for aggregate output, industrial production, and aggregate investment all demarcate distinct epochs of relative growth and stagnation.

Table 3 presents the average effects of cyclical downturns, averaging the data from columns (3) and (4) in Table 1, for crisis and expansion periods defined by this dating system. There are sharp and consistent differences between the crisis and expansion periods by these measures as well.<sup>3</sup>

<sup>2</sup>Since we do not have comparable data for the years before 1890, we cannot be certain, of course, that the cycle beginning with the peak in 1890 was the first nonreproductive cycle in this particular sequence.

<sup>3</sup>Some readers may question our use of the NBER dating of peaks and troughs rather than a cycle dating by peaks and troughs in the unemployment rate. We prefer our choice since we are concerned with the effects of downturns in business activity in general, not simply with movements in unemployment; there are two

We note, finally, that our dating identifies crisis periods which correspond to eras in which historians have identified intense institutional conflict and innovation. What have been termed "key" presidential elections, those of 1892 and 1896, and of 1932 and 1936, fall within our first two crisis periods. All three crisis periods, defined by our method, have witnessed intensified class conflict and sharp debates over major economic policy issues.

We conclude that this empirical evidence, however provisional, establishes strong initial support for our model of the relationship between the business cycle and long swings.

---

"perverse" business cycle downturns, indeed, which did not result in increases in the unemployment rate. Our results are robust if we use an analogous unemployment dating scheme, in any case: the numbers corresponding to the averages in column (1) of Table 3 are +2.66, +0.94, +1.23, +1.55, -1.93, -0.98, and -1.49 percent, respectively.

#### REFERENCES

- Bowles, Samuel, "Competitive Wage Determination and Involuntary Unemployment," mimeo., University of Massachusetts-Amherst, 1981.
- \_\_\_\_\_, and Gintis, Herbert, "The Crisis of Liberal Democratic Capitalism: The Case of the United States," *Politics and Society*, Winter 1982, 11, 51-92.
- \_\_\_\_\_, Gordon, David M., and Weisskopf, Thomas E., *Beyond the Waste Land: A Democratic Alternative to Economic Decline*, New York: Doubleday Books, 1983.
- Gordon, David M., "Stages of Accumulation and Long Economic Cycles," in T. Hopkins and I. Wallerstein, eds., *Processes of the World-System*, Beverly Hills: Sage Publications, 1980.
- \_\_\_\_\_, Edwards, Richard, and Reich, Michael, *Segmented Work, Divided Workers: The Historical Transformation of Labor in the United States*, New York: Cambridge University Press, 1982.
- Schor, Juliet B., and Bowles, Samuel, "The Social Wage and the Labor Process: Measuring Some Influences on Worker Resistance," mimeo., Williams College, 1982.
- Weisskopf, Thomas E., "Marxian Crisis Theory and the Rate of Profit in the Postwar U.S. Economy," *Cambridge Journal of Economics*, December 1979, 3, 341-78.
- \_\_\_\_\_, Bowles, Samuel, and Gordon, David M., "Hearts and Minds: A Social Model of Aggregate Productivity Growth in the United States, 1948-1979," mimeo., Economics Institute of the Center for Democratic Alternatives, 1983.

## THE CHANGING FORTUNES OF REGIONS

### On the Effects of Federal Aid

By GEORGE TOLLEY, RONALD KRUMM, AND JEFFREY SANDERS\*

Concern with the effects of aid by higher units of government to cities and other localities has ranged from enumeration of geographical distribution of aid (see, for example, T. L. Muller, 1982) to various attempts to identify and take into consideration local reactions that influence the ultimate effects. The latter include studies that focus on particular programs (see, for example, the analyses included in N. J. Glickman, 1980), effects on local government expenditures (see, for example, R. C. Fisher, 1982; P. N. Courant et al., 1979; and M. B. Johnson, 1979), and broader issues such as the concern with urban decline in many areas (see, for example, K. L. Bradbury et al., 1981, 1982). In this paper we attempt to contribute to the analysis of intergovernmental aids and other federal programs by suggesting a general framework for analyzing their effects and utilizing the framework to estimate some of the impacts.

The framework concerns first the values of aid to localities and persons as affected by in-kind restrictions. Based on the values of aid, the geographic redistributions of income among areas due to all federal actions are considered, requiring estimation of the geographic distribution of nonaid items including taxes, place of federal purchases of goods and services, and the location of benefits of the purchases—along with attendant interregional multiplier effects due to induced changes in demand for local goods and services.

In a very short run with no mobility of factors of production, the foregoing geographic redistributions would translate di-

rectly into changes in per capita incomes among areas. At another extreme, in a constant cost world with infinite elasticity of supply of all factors of production to any location, the geographic redistributions due to federal actions would have no welfare significance, since factor flows would keep per unit remuneration of each factor everywhere the same. Even in the long run, costs are not invariant with scale of activity, and there are location-specific factors of production including land. The framework offered here brings in cost and production function analysis along with dynamic factor mobility considerations to get at the welfare implications of the geographic redistributions due to federal actions.

#### I. A Framework for Analyzing Grants to Localities

Consider local government provision of a good or service equal to  $q_0$  which is subject to a categorical aid matching requirement equal to  $q_M$ . Suppose  $q_M > q_0$ , so that a community choosing to receive aid equal to  $a$  must substitute out of all other goods and services until  $q_M$  is provided. In this situation, the budget constraint for the community is shifted out for values of  $q > q_M$ , but is discontinuous between  $q_M$  and  $q_M + a$ . Choice of  $q > q_M$  could come about as a result of the income effect on the community if the income elasticity of demand for the good in question was sufficiently high. This would happen only if

$$(1) \quad \eta_q > [1 + (q_M - q_0)/a]/W_q,$$

where  $\eta_q$  is the income elasticity of demand, and  $W_q$  is the proportion of expenditures devoted to the aided good. The effect of the grant would be the same as if it were a pure income transfer to the community. If the

\*University of Chicago. Partial financial support from the National Science Foundation is gratefully acknowledged. This paper provides a summary of some of the findings of our earlier 1982 study.

TABLE 1—VALUE TO COMMUNITY OF A DOLLAR OF CATEGORICAL AID ( $G/a$ )

		Demand Elasticity for Aided Good				
$q_0/a$		-.1	-.2	-.5	-1.	-2.
$q_M/a = .5$	.10	-.389	-.375	-.309	-.129	.175
	.25	-.222	-.188	-.042	.198	.475
	.40	-.056	-.001	.193	.429	.651
	.45	.000	.062	.265	.492	.693
$q_M/a = 1.0$	.05	-.944	-.937	-.898	-.766	-.418
	.20	-.778	-.750	-.620	-.331	.065
	.50	-.444	-.375	-.125	.193	.500
	.90	.000	.116	.395	.618	.783
$q_M/a = 3.0$	.95	.065	.175	.449	.657	.807
	.50	-2.444	-2.375	-2.063	-1.460	-.672
	1.0	-1.889	-1.751	-1.250	-.613	-.000
	2.0	-.778	-.531	.000	.386	.657
Match is Less than Expenditure without Program ( $q_M < q_0$ )	2.5	-.226	.030	.438	.675	.825
	2.9	.204	.425	.698	.833	.912
	.05	.005	.012	.048	.152	.358
	.10	.011	.025	.091	.240	.463
	.50	.050	.123	.333	.544	.732
	1.0	.100	.234	.500	.639	.812
	2.0	.197	.401	.667	.811	.899
	5.0	.419	.647	.833	.912	.938
	10.0	.614	.792	.909	.953	.976
	$\infty$	1.00	1.00	1.00	1.00	1.00

aided good accounts for only 5 percent of total expenditures in the community, however,  $\eta_q$  would have to exceed 20 for the community to choose  $q > q_M + a$ , rather than  $q = q_M + a$ , a circumstance which seems highly unlikely.

When the matching requirement is lower than expenditure in absence of program ( $q_M < q_0$ ), the community cannot lose by participating in the program. The result where the effect of the aid is equivalent to an unconstrained transfer will be obtained if the income elasticity condition in equation (1) holds. With  $q_M < q_0$ , the numerator on the right-hand side of equation (1) is less than unity, making the condition more conceivable but still highly unlikely if aid is high relative to expenditures in absence of the program.

The analysis helps explain why categorical aid programs are likely to result in more expenditures on aided goods than would be chosen if the same amount were given in unconstrained form. Finding the perceived value of the aid to the community requires

bringing in price elasticity considerations. For  $q_M > q_0$ , and with constant elasticity of demand equal to  $-1/b$ , so that the inverse demand curve is  $f(q) = p_q(q/q_0)^{-b}$ , the perceived value to the community is

(2)

$$G = \int_{q_0}^{q_M+a} p_q(q/q_0)^{-b} dq + p_q(q_0 - q_M).$$

Table 1 illustrates how  $G/a$ , the perceived value to the community per dollar of aid, varies according to combinations of  $b$ ,  $q_M/a$ , and  $q_0/a$ . Of interest are the entries where  $G/a$  is calculated to be negative, indicating that the community is not expected to participate in the program unless other considerations (for example, income distribution) are dominant in the decision-making process. Participation is expected only if matching requirement per dollar of aid is low, particularly for demand elasticities of unity or less. Similar calculations of  $G/a$  for the case when  $q_M < q_0$  are shown in the lower portion of Table 1.

The values of  $G/a$  in Table 1 are generally far from one. A conclusion is that for matching requirements most frequently encountered, the value of a dollar of categorical aid can be worth substantially less than a dollar to the community.

## II. Regional Expenditure Calculations

Once the value of in-kind aid is determined according to the considerations in the preceding section, spatial effects of government actions can be ascertained beginning with federal government expenditure and revenue data by locality. The concern in this section is with determination of net income to communities resulting from federal government aids and more generally, overall tax and expenditure policies. For such purposes, seven expenditure categories are considered as they differ by region: welfare program cash transfers to individuals; in-kind transfers to individuals; grants to local governments and agencies; retirement-related income transfers to individuals; defense procurement expenditures; defense salaries; and purchases of other goods and services.

The value of welfare program cash transfers and retirement-related transfers are assumed to have a dollar-for-dollar value to recipients in the area, because there are no restrictions on how the income is spent. On the hypothesis that the diminution in value due to restrictions associated with in-kind transfers to individuals is ordinarily small, these are also assumed to have a dollar-for-dollar value. The value of aid to localities appears to be importantly subject to diminution in value due to matching requirements and restrictions on how funds are spent, and therefore the federal dollars expended in this category are reduced by a suggestive value of 25 percent.

Total federal taxes paid are calculated to be the regional share of all individual and corporate taxes paid in the region relative to the national total, multiplied by the sum of all federal expenditures. The incidence assumption for corporate taxes, to be refined, is that they are spread to all capital approximated by the distribution of population. From these calculations, the costs of

each program were estimated by prorating the total federal outlay in the nation on the programs to the regions, according to the region's proportion of total revenue collected. These were then subtracted from actual expenditures resulting in a net expenditure by program type in each region.

Much income received in each region will be spent by recipients in the region where received. If  $m$  of each dollar received is expended on locally produced goods and services, the additional income generated in the region is given by  $m/(1-m)$ . It is assumed here that  $m = .6$ , so the additional income generated is equal to 1.5 times initial net income received. Further interregional input-output feedbacks could be modeled.

The effects of the remainder of federal expenditures, namely federal purchases of goods and services, and salaries, enter into the analysis in two ways, first with regard to the locations where benefits of government output are received and paid for, and second, where the purchase of the goods and services take place. Defense benefits are assumed to be ubiquitous to the nation and also assumed to have no differential regional income effects. Benefits of nondefense purchases of goods and services are allocated among states in proportion to population. Taxes paid for these services were then subtracted using the same allocation of a dollar of federal taxes among regions discussed above, and the same multiplier was applied to arrive at the regional income induced impact.

With regard to the second issue of effects of where government purchases take place, estimates are not available beyond the point of disbursement of the federal government. It was assumed that .7 of each dollar dispersed in a region was directly spent in that region and that of the .3 left over, the direct expenditure in the region was equal to its population share. The latter is justified on the ground that the myriad of services employed in production of final government goods and services is proportional to the population in an area. The total cost to the nation of these federal purchases was allocated among regions according to the taxes paid as an estimate of what private spending

TABLE 2—REGIONAL INCOME IMPACTS OF FEDERAL EXPENDITURE AND TAX POLICY  
(Millions of Dollars)

North East	Mid- Atlantic	East North Central	West North Central	South Atlantic	East South Central	West South Central	Pacific	Mountain
3,456	-36,716	-62,990	3	54,693	23,079	584	6,394	11,497

patterns would otherwise have been. The net income was then subjected to the same multiplier as for other expenditure categories to obtain a regional income induced impact.

The sum of net incomes generated by each category based on the above procedures is given in Table 2. The results are to be contrasted with simple unadjusted calculations of federal tax collections and federal expenditures by states reported in the press. Table 2 attempts to give more accurate estimates of the extent to which communities in the Mid-Atlantic and East North Central regions lost to income on net, while those in other regions gain income due to the federal tax and expenditure policies.

### III. Determination of Beneficiaries and Adjustment

There will be a strong if not complete tendency for aids to individuals and retirement payments to go to designated beneficiaries. A more difficult issue is who are the beneficiaries of other transfer categories, namely aids to localities. One possible set of beneficiaries consists of owners of land and existing capital stocks whose returns are affected because of the willingness of consumers to pay for access to the services provided. Also, local governments may have their finances affected by the conditions of the aid itself and the resulting local multiplier impacts. An example which illustrates the extent to which benefits go to the first set of beneficiaries is aid to housing. Consider the case where the government supplies housing in an area equal to  $Q$  of the total quantity privately supplied. The benefits to consumers as a fraction of value of consumption is

$$[Q/(b+c)][1+.5Q/(1+c/b)]-S,$$

where  $b$  is the absolute value of the elasticity of demand,  $c$  is the elasticity of supply, and the queued subsidy as a fraction of the total market value of housing is  $S$ . The greater is the sum  $b+c$ , the fewer are the benefits passed on to consumers, with the main effect of the program being to replace private supply. The addition to the percent of total market value of sites where the new supply of the good is sold is  $Q[1-Q/(b+c)]-S^1$ , where  $S^1$  is the subsidy plus expense borne by the seller as a proportion of total market value. The greater is the sum  $b+c$ , the greater is the increase in site values. A similar expression can be developed for sites where the government supplied good is not sold. The conjecture made here is that for reasonable values of  $b$  and  $c$ , while benefits to consumers are not negligible, a goodly portion of benefits take the form of increases in land values at sites where this government supplied good is sold.

A broader issue is concerned with interregional adjustments to expenditure and tax policies associated with the multiplicative effects on employment and returns to land and other factors of production. In a national economy with perfect mobility of persons between regions, disparities in tax and expenditure policies may lead to little, if any, change in real wages received by workers, regardless of location. Although extensive development of the implications of such adjustments is beyond the scope of this paper, one conjecture is that in a full-employment economy, the redistributions of income from one place to another would only change the location of employment of persons, not necessarily their real income. Similarly, the provision of benefits of services that exceed their costs in a locality may only result in a change in the nominal wage received in the area, leaving workers as well off as before. Non-

workers, too, will adjust to differences in payments and cost of living. Given these longer-run adjustments, the differences in incomes shown in Table 2 will reflect mainly shifts in location of factors or production with little or no change in real remuneration to mobile factors. Long-run changes in rates of remuneration are then limited mostly to owners of land or other locationally fixed factors accounting for a relatively small part of total income. As suggested in our earlier paper (1982), as much as one-tenth of the locally generated or lost income may go to land. The long-run income redistributions to location-specific factors would then be estimated by multiplying each of the numbers in Table 2 by one-tenth. For example, the conclusion would be that the long-run income redistributions consist mostly of raising land values in the South Atlantic by \$5.47 billion, lowering them by \$6.30 billion in the East North Central, and similarly for the other entries in the table.

Noninfinite elasticities of labor supply associated with life cycle migration between regions, taste differences, and the presence of structural unemployment would alter the estimates just given with others besides claimants to land income benefiting from positive net government expenditures in a region, an outcome which is likely in a shorter run. A labor supply stock adjustment model distinguishing between expansions and contractions is needed to explain the time path from short to long run.

Another type of effect is the impact of federal tax and expenditure policies on the fiscal positions of local governments as the cost of supplying local services changes by a different amount than revenues collected. For nonconstant cost supply of these services (for example, through fixed costs of infrastructure), the induced scale effects may alter local government fiscal positions, depending on whether or not the area is growing or declining. For growing areas, both fixed and variable costs may be incurred, while in declining areas, only variable costs will change. Our more extended study suggests that these effects are less than the above regional redis-

tributions, but still nonnegligible and exacerbated by lags in local government behavior.

#### IV. Conclusion

The numbers presented here for the effects of all federal actions invite refinement. The framework should be applied to individual programs. The analysis brings out the need for empirical attention to the relation between short- and long-run factor adjustments if regional effects of federal programs are to be more fully understood.

#### REFERENCES

- Bradbury, K. L. et al., *Futures for a Declining City: Simulations for the Cleveland Area*, New York: Academic Press, 1981.
- , *Urban Decline and the Future of American Cities*, Washington: American Enterprise Institute, 1982.
- Courant, P. N. et al., "The Stimulative Effects of Intergovernmental Grants: Or Why Money Sticks Where It Hits," in P. Mieszkowski and W. Oakland, eds., *Fiscal Federalism and Grants-in-Aid*, Washington: The Urban Institute, 1979.
- Fisher, R. C., "Income and Grant Effects on Local Expenditure: The Flypaper Effect and Other Difficulties," *Journal of Urban Economics*, November 1982, 12, 324-45.
- Glickman, N. J., *The Urban Impacts of Federal Policies*, Baltimore: The Johns Hopkins University Press, 1980.
- Johnson, M. B., "Community Income, Intergovernmental Grants, and Local School District Fiscal Behavior," in P. Mieszkowski and W. Oakland, eds., *Fiscal Federalism and Grants-in-Aid*, Washington: The Urban Institute, 1979.
- Muller, T. L., "Regional Impacts," in J. L. Palmer and I. V. Sawhill, eds., *The Reagan Experiment*, Washington: The Urban Institute, 1982.
- Tolley, G. S., Krumm, R. J. and Sanders, J., "Spatial Impacts of Federal Expenditure and Tax Policy," unpublished manuscript, University of Chicago, 1982.

# Industrial Bases and City Sizes

By J. VERNON HENDERSON\*

How do industrial bases vary across cities and do they vary by city size? Are smaller types of cities oriented towards heavy manufacturing or towards traditional service activities? What size cities do high-tech and modern service industries gravitate towards? Why are these questions relevant?

Over the last thirty years, there has been a relative shift in national production patterns away from heavy manufacturing and blue-collar occupations, and towards professional activities, communications, finance, insurance, real estate, and high-tech activities. Should we expect these gains and losses to be spread across cities, or are certain cities likely to be gainers and others losers, in terms of population? Are losers likely to experience permanent declines in population; or, for some, will their losses level off or will they rebound?

If the growth industries of an economy operate best in larger cities while declining industries tend to reside in smaller cities, this implies national population will shift from smaller cities to locating in larger cities. This increase in urban concentration means there will be changes in other patterns of resource usage. Efficient levels of per capita resources devoted to urban infrastructure (roads, sewers, parks, etc.) tend to rise with city size, and it is a common perception that the magnitudes of certain consumer externalities (congestion, air and water pollution, crime, etc.) rise with city size. This has implications for public policies governing the allocation of public investment and governing the regulation of externalities. Moreover, the changes in population allocation between larger and smaller cities, itself, involves to some extent a painful process of population upheavals and there may be public policies which can alleviate the costs of the adjustment process.

In this paper, I outline a conceptual framework and empirical methodologies from which some of these issues can be examined.

## I. A Conceptual Framework

Since the concern here is with population allocation across cities, it is necessary to have a model of a system of cities as a framework for posing questions and determining methodologies for answering those questions. I start by outlining some basic notions and presenting some basic facts about the system of cities in the United States.

As suggested in Edwin Mills (1967), large cities form because there are scale economies in production which lead workers and firms to cluster in close spatial proximity in large agglomerations. At the same time, there are consumption diseconomies connected with people clustering in urban areas, such as commuting cost increases in a monocentric city, which eventually offset the production scale benefits at the margin as city size increases, limiting cities to various equilibrium sizes. Determining how city sizes are limited requires an analysis of how new cities form and compete with each other. In the United States, we presume competitive processes where autonomous local governments compete for residents, firms compete in factor and output markets, and developers compete for migrants and investment with new urban developments.

Given mechanisms for competition among cities, we might expect cities to specialize in the production of one traded good or set of interrelated goods, recognizing that the extent of any specialization is limited by the costs of trading goods across cities, so that in fact most goods such as housing and retail and personal services are nontraded across cities. There are several explanations of why there is specialization among traded goods, and I focus on one—the nature of scale economies. To be consistent with perfect

\*Brown University. Financial support from the National Science Foundation, grant numbers SOC 79-01592 and SES 80-13482, is gratefully acknowledged.

competition, scale economies are modeled as being parametric economies of scale external to firms. They have two characterizations.

First, they may be economies of *localization*, internal to each industry, where scale is measured by total employment (or output) in *that* industry in *that* urban area. Scale could reflect Adam Smith economies of intra-industry specialization, labor market economies for workers with industry-specific training, or scale of "communications" among firms affecting the speed of adoption of new innovations. Second, they could be economies of *urbanization*, external to the specific industry, and resulting from the level of all economic activity internal to a city, measured by, say, total city population. In this case, only the size of the city, not its industry composition, affects the extent of scale effects relevant to a particular industry.

If scale effects are ones of localization, then for a *given* city size and associated cost of living, scale effects and hence incomes are maximized by concentrating local export employment all in one industry, rather than dissipating the scale effects by spreading employment over many industries. However, if scale effects are ones of urbanization, then this specialization may not matter since it is the general level of economic activity rather than its industry specific concentration which enhances productivity.

Empirical work on the United States (see my 1982b paper) strongly supports the hypothesis that, for manufacturing export industries, economies of scale are ones of localization, not urbanization. Amongst all two-digit industries, only nonmetallic minerals has significant positive urbanization economies. This would suggest that there should be different types of cities specialized in the production of the different manufacturing industries exhibiting localization economies. For the 243 SMSA's in 1970 using cluster analysis on patterns of per capita employment with a strong criterion, I found 6 textile cities, 5 food processing cities, 4 communication equipment cities, 6 pulp and paper cities, 7 apparel cities, 10 steel cities, 3 leather product cities, 4 petrochemical cities, 5 industrial machinery cities, 4 shipbuilding cities, 6 aircraft cities, and 12 auto cities.

Because different production activities involve different degrees of scale economies, different types of cities will have different efficient sizes. These notions are critical.

First, there is a direct link between national production patterns and numbers of different types of cities producing the goods composing national output. Second, because different types of cities have different sizes, there is a direct link between national output patterns and the size composition of cities. If national output patterns shift in favor of goods produced in larger types of cities, then urban concentration will increase. This shifting process will involve some smaller types of cities changing production patterns and growing into larger types of cities, as well as some smaller types of cities steadily evaporating over time.

There remains the question of why a particular city specializes in one particular good, rather than another. Suppose there is a set of urban sites in an economy upon which cities can form. Each site has a natural endowment from the set of site characteristics—access to various natural resource deposits, access to coastal ports, climate, terrain, altitude, etc. Industries compete for sites resulting in an equilibrium allocation where, for example, steel-type cities go to sites with access to limestone, coal and iron ore, while large international market-oriented metro areas go to coastal ports and sites with low endowments of all characteristics may remain unoccupied. Sites with high levels of all attributes go to the the largest types of cities (who can bid the most for them). In terms of shifting national production patterns, for cities specialized in declining industries, those cities whose site amenities can also be used by growing industries will shift production patterns. Among the remaining cities, those with the best site amenities for the declining industries will remain producing that type of output, and those with the worst site amenities will die out over time.

This characterization of an economy composed of different types of cities specialized in different activities has limitations, because of the transport costs of trade across cities and the probable existence of urbanization economies for certain nonmanufacturing in-

dustries. First, transport cost considerations present a problem in the geographic interpretation of patterns of specialization. My model to date assumes cities are basically monocentric. If all of our industries were footloose, different types of "monocentric" cities would cluster together to reduce the transport costs of intercity trade. By doing so, they might *geographically* form a large multinucleated metropolitan area consisting of a cluster of *economic* "cities."

What gives us geographically monocentric small or medium size urban areas specialized in production are natural resource considerations (see my 1982a paper). Cities spread out spatially so as to have access to raw materials. Thus, spatially distinct specialized cities are those involved in weight-reducing, resource-using industrial production such as primary metals, heavy machinery, wood products, and some food processing to pick obvious examples, assuming economies are ones of localization. On the other hand, even if all economies were localization ones, footloose industries such as financial and business services and light and high-tech manufacturing might cluster together to form large multinucleated urban areas (perhaps clustering around large types of resource using cities, see my 1982a paper). There is also the notion that footloose industries such as services and market-oriented light manufacturing may experience urbanization economies.

While half of the 243 SMSA's in 1970 might be classified as specialized in one manufacturing industry, there are many diversified metropolitan areas engaged in a variety of service-oriented activities, and some engaged in diverse manufacturing activities. Remaining SMSA's are government towns (state capitals and state university cities) or cities servicing rural areas in a traditional central place model (with employment in wholesaling, warehousing, and transport services).

In summary, to predict spatial shifts in urban concentration, we must be able to relate changes in national production patterns to changes in the need for different types of cities and to identify whether growth industries will be industries which cities spe-

cialize in or industries with strong urbanization economies which will be attracted to existing large metropolitan areas. In addition, to determine which existing cities will be winners versus losers, we must be able to match amenity needs of industries with site characteristics of cities. Finally, we must be able to isolate the additional forces of technological change which may tend to alter all city sizes. For example, technological improvements in communications may tend to weaken the benefits of agglomerating production activity together, having a negative impact on all city sizes.

## II. Methodology

I examine two ways to start making predictions about the impact of changes in national output patterns upon the allocation of population to different sizes and types of cities. The back-of-the-envelope method starts by looking at gross patterns of city sizes versus employment in gross industrial categories, to see how local employment in any one industry varies with city size. The second method looks at detailed industries and estimates micro models of industrial location, where the probability of finding an industry in a city and the size of the industry in cities where it is found is a function of price and amenity variables.

### A. Gross Patterns

To see how production patterns generally vary with city size, I regress the *share* of employment in each SMSA for different gross industrial categories on urban population (*POP*) and its square (*POPSQ*), as well as regional dummies and a dummy variable for SMSA's with more than 15 percent of their employment in state or federal public administration. The gross industrial categories are resource bound manufacturing (primary metals, machinery, autos, ships, rail, wood products, and some agricultural processing), footloose manufacturing (everything except resource bound and high-tech), high-tech manufacturing (aircraft, computers, instruments, weapons, medical equipment), professional services (health, legal, engineering,

TABLE 1—EMPLOYMENT SHARES

	Resource Bound Manu.	Foot Loose Manu.	High- Tech Manu.	Prof. Serv.	Whole- sale	Bus. Serv.	FIRE
POP	$-.127 \times 10^{-4}$ (1.26)	$.145 \times 10^{-4}$ (1.64)	$.841 \times 10^{-5}$ (2.61)	$.205 \times 10^{-5}$ (.91)	$.571 \times 10^{-5}$ (2.56)	$.539 \times 10^{-5}$ (5.73)	$.855 \times 10^{-5}$ (4.23)
POP SQ	$.408 \times 10^{-8}$ (.35)	$-.141 \times 10^{-8}$ (1.39)	$-.841 \times 10^{-9}$ (2.28)	$-.571 \times 10^{-9}$ (.22)	$-.487 \times 10^{-9}$ (1.91)	$-.349 \times 10^{-9}$ (3.25)	$-.474 \times 10^{-9}$ (2.50)
REG NC	.050 (3.61)	-.101 (8.36)	-.012 (2.70)	.001 (.41)	.005 (1.53)	-.002 (1.67)	-.001 (.24)
REG S	-.061 (4.58)	-.092 (7.90)	-.012 (2.72)	-.003 (1.10)	.014 (4.92)	0 (.06)	.005 (1.91)
REG W	-.038 (2.37)	-.163 (11.64)	-.006 (1.15)	.004 (1.22)	.009 (2.57)	.004 (2.93)	.006 (1.82)
Govt. Dummy	-.041 (3.15)	-.037 (3.30)	-.003 (.77)	.013 (4.49)	-.018 (6.14)	.001 (.79)	$-.23 \times 10^{-3}$ (.09)
Constant	.176	.191	.023	.080	.049	.012	.042
R <sup>2</sup>	.31	.41	.08	.10	.21	.26	.16
Percent $\Delta$ Share/ Percent $\Delta$ POP	-.029	-.064	.223	.012	.046	.210	.084
Mean Share	.15	.10	.02	.08	.06	.02	.05
Pop. for Max. Share (in mil.)	n.a.	5.14	5.00	17.95	5.87	7.7	9.02

and architectural, accounting, auditing, and bookkeeping and miscellaneous), wholesale services (wholesale, trucking, and warehousing and storage), business services (excluding repairs), and "FIRE" (finance, insurance, and real estate). The data is from the Sixth Count of the 1970 Population Census for 242 SMSA's.

In Table 1, the share of resource bound manufacturing appears to decline (weakly) with city size, while all other activities rise (except possibly professional services). Other manufacturing activities tend to peak at SMSA sizes of around 5 million, while the expanding white-collar sectors of the national economy—business services and FIRE—tend to peak at 8–9 million. Given its fast rate of increase, high-tech seems to be very concentrated in larger SMSA's. Note the general hypothesis that footloose activity, relative to resource bound activity, will cluster in large SMSA's is borne out. In terms of regions, manufacturing (except for resource bound in the NC) is strongly oriented to the NE, while the expanding white collar sectors are oriented to the W. These should reflect regional differentials in production or consumption amenities (climate for white-collar

workers). Manufacturing tends to avoid SMSA's with government workers, presumably because of the low level of scale externalities relevant to manufacturing (given the concentration of government).

These results are consistent with other casual impressions. Urban concentration as measured by, say, the percent of the national urban population in urban areas over 1 million, increased from 39 to 47 percent between 1950 and 1970, and then leveled off. This corresponds to the shift out of resource-oriented manufacturing into other activities. Similarly, the regional shifts to the W and S are consistent with the amenities relevant to service and white collar activities versus those relevant to the declining manufacturing industries of the NE and NC.

### B. Micro Models

To determine which particular cities are going to be winners vs. losers, in terms either of attracting nationally growing industries in an economy or of retaining, at current employment levels, nationally declining industries, a starting point is to estimate models of whether a particular industry is found

in a particular urban area and, if so, what its scale (output or employment) is. The determinants of these outcomes are local price and amenity variables.

Employment in a particular industry occurs in urban area  $i$  if profits exceed some critical level. Profits are  $\pi_i = F(X_i) - \phi_i$ , where  $X_i$  is a vector of input prices, amenities, local infrastructure services, access measures to national markets, scale effects, etc. The  $F(\cdot)$  is the profit function to be estimated and  $\phi_i$  a random variable representing unknown location specific effects. Incorporating the critical level of profits into the  $F(\cdot)$  function, an industry is found in an urban area if  $F(X_i) - \phi_i \geq 0$ . Using a standard probit specification, the probability  $p_i$ , of finding this particular industry in urban area  $i$  is

$$(1) \quad p_i = \int_{-\infty}^{F(X_i)} g(\phi) d\phi,$$

where  $F(X_i)$  is the "long-run" profit function. If annual employment,  $L_i$ , in that industry is based on this profit function, where  $q_i$  is the price of labor,

$$(2) \quad L_i = -\partial F(X_i) / \partial q_i.$$

In this case, (1) and (2) can be estimated jointly by maximum likelihood methods and the parameters of  $F(X_i)$  defined in absolute terms.

I examined four growth industries—insurance (exclusive of sales), investment, aircraft, and computers. Due to censorship, Census of Manufacturing data cannot be used to estimate equation (1). The other data I had was detailed employment data (for 229 industrial categories) from the 1970 Population Census. Unfortunately, due to survey problems of self-classification, our industries registered employment in almost all SMSA's—even for aircraft and computers only 8–9 percent of all SMSA's had zero employment. This is really not enough to estimate the discrete choice model in equation (1).

Although we can get estimates (albeit biased ones) of equation (2) alone, the interpretation of the results is difficult unless we also have estimates of  $F(X_i)$  in equation (1). For example, an argument in equation (2) is

scale as measured by urban population. Without knowing  $F(X_i)$ , we cannot determine whether the scale coefficient represents the general (neutral) impact of urban scale on profits or a measure of the bias (nonneutrality) of scale effects for relative labor usage. Moreover, given the roughness of the employment measures, where, for example, by self-identification employees in repairing and servicing of aircraft may be classified in manufacturing, equation (2) scale variables may include local demand effects for services. The solution is to have data for detailed solely export industries. Unfortunately, due to censorship of the Industrial Census, such data may not be available in the United States (although I have it for Brazil).

There is not space to formally report the results I did obtain for equation (2) for the four growth industries. In general, wages had an anticipated negative (own price) impact and the price of capital a positive impact (substitutability with labor) on local employment. For insurance (percent female labor force participation), investment (percent adults with 4+ years college), and computers (percent manufacturing workers with 4+ years high school), measures of labor market conditions were important amenities. For aircraft, employment in an SMSA was associated with fewer heating degree days. Finally, for aircraft and computers, the level of other high-tech employment positively affected own local employment. These results suggest the methodology is useful, but better data is needed.

## REFERENCES

- Henderson, J. V., (1982a) "The Impact of Government Policies on Urban Concentration," *Journal of Urban Economics*, November 1982, 12, 280–303.
- , (1982b) "Efficiency of Resource Usage and City Size," Working Paper No. 82–14, Brown University, 1982.
- Mills, E. S., "An Aggregative Model of Resource Allocation in a Metropolitan Area," *American Economic Review Proceedings*, May 1967, 57, 197–210.

# Economists, Economics, and State Economic Policy

By ROGER J. VAUGHAN\*

In the past five years, state governments have experimented with a growing number of policies to stimulate local development. Tax incentives, business loan and loan guarantee programs, and regulatory reforms have proliferated as states seek to improve their business climates and create jobs. Public action has been prodded by plant closings, interstate differences in economic growth, the availability of discretionary federal dollars, and by the urgings of the business community anxious to tilt the state cornucopia their way. By and large, these efforts have been undertaken without the results of academic research and advice.

In spite of the advances in the science of economics during the last two decades, the ability of academic economists to assist in the formulation of state economic development policy is very limited. When they do offer advice, it is usually couched in vague terms and largely ignores the institutional and legal framework within which decisions must be made. Although most economists typically feel that no action is the preferred course, this attitude ignores the fact that, in many instances, an event such as a plant closing has generated sufficient public attention to a problem that action must be taken regardless of the economic merits. Yet economic analysis can offer important insights that could aid state economic policy. This paper describes the barriers that prevent closer cooperation between policymakers and economists at the state level. It also outlines areas where academic research could be directed to provide immediate assistance to state policymakers.

Most economists are unfamiliar with state government. Many years of conducting federally sponsored research and often of serving for some years in various federal agencies have established close links between academia and federal agencies. Formal links be-

tween universities and state agencies have not been established. In spite of the significant role that state revenues play in underwriting much of the university system, state governments—neither the line agencies nor the governors' offices have used their financial leverage to establish strong economic research ties with state campuses.

A second reason for the lack of familiarity is that there are no economic textbooks that could be used to introduce faculty and students to state policymaking. Differences among states with respect to constitutional and legislative powers have led to a bewildering variety of tax and regulatory structures that must be taken into account when designing economic policy. The best textbooks aim at a national market and therefore have focused on federal tax and regulatory policy. Until there are adequate books available, it is unlikely that there will be any significant improvement in academic awareness of state issues.

Few graduate programs in economics include courses in public policy that provide the students with the necessary research tools to understand the institutional and legal framework within which policy decisions must be made. At the very least, students should be aware of how legislation is enacted as well as how to use a law library to determine the powers granted to different levels of government by the state constitution and the current status of tax and regulatory policy. But to teach these courses requires experience and skills possessed by very few academic economists and would involve reform of the curricula. Although these changes cannot easily be made, they are needed if economics departments are to work more closely with state governments.

Finally, the way in which economics is taught in most departments is unsuitable for analysis of many local economic policy issues. The problems that state legislatures and bureaucrats face are usually the problems associated with economic adjustment.

\*Senior fellow, Gallatin Institute.

Changes in world trade patterns, in consumers' tastes, in resource prices, or in technology lead to rapid declines in some industries and areas, which are only indirectly offset by growth in other areas. The expansion of the microprocessor industry in California has done little to compensate, directly, the steel workers rendered jobless in Lackawanna, New York. By failing to move beyond a comparative static economic framework, teachers do not provide the necessary analytic tools needed to deal with state policy issues. In confronting the problem of what state actions may be needed to help the local economy adjust to change, policymakers must wrestle not only with diagnosing what needs to be done to make the economy function more efficiently, but also with devising a set of policies that distributes the cost of economic adjustment in a politically acceptable way. Economic analysis of the costs and consequences of alternative policies (including the distribution of these consequences) would be an invaluable tool for state policymakers. The power of special interest groups to manipulate the political process to exact preferential treatment at the expense of a larger, but less-informed community has often been lamented. Yet injecting more information into the process will require a greater sensitivity by academic economists to politics and policies at the state level.

Since state policy has received so little attention, research needs are almost unlimited. However, in response to recent changes in the direction of state economic strategies, three areas would be immediately productive: capital markets; state regulatory policy; and state tax policy.

*Capital Markets.* Economists typically view capital markets as complying with most of the requirements of a perfectly competitive environment. Yet the complex effects of federal and state taxes and the detailed regulation of financial institutions by both state and federal agencies probably distort flows of capital between different types of investment (differentiated by risk and liquidity). The restrictions imposed on the investment of public pension funds, the asymmetric treatment of capital gains by both state and

federal tax codes, the regulation of insurance companies' portfolios, and other state actions may lead to distortions in the allocation of credit that impose special problems on certain types of companies. Yet there is little known about the magnitude or even the direction of these effects. In 1938, William O. Douglas, then Chairman of the Securities and Exchange Commission, pointed out that no one had studied the regional operation of capital markets. That gap has yet to be filled. Recent surveys conducted by the Federal Reserve Board as part of a research project concerned with the credit problems of small business are a potentially rich data source.

*State Regulatory Policy.* The political climate in Washington has precipitated a growing volume of research concerned with the effects of federal regulatory policy on economic efficiency and income distribution. Although many states are currently reviewing their own regulatory practices, they do not have a comparable body of research to assist them in this task. With the rapid evolution of new technologies and their application to business activities, many state regulatory policies have been rendered obsolete. For example, in many states, the right of municipalities to grant sole franchises to cable TV corporations includes the right to exclude "competing" modes of communication. Yet the dramatic reduction in the cost of master antennae systems, the evolution of low-powered TV and other rivals to cable have probably eroded any monopoly power once enjoyed by cable systems. Yet the economics of the communications industry has received scant attention, and there is little or no empirical evidence concerning the competition between alternative modes that might assuage the concerns of consumer-oriented legislators. The breakup of AT&T will lead to considerable confusion among state utility-regulating agencies without the results of careful research. The massive regulation of financial institutions, of university-based research, of occupational accreditation, and of energy generation is also being challenged by rapidly changing technologies. States are ill-equipped to meet these challenges, especially as budget cutbacks lead to sharp cuts in "non-essential" functions such as research.

*State Tax Policy.* In the wake of Proposition 13 in California, many states undertook extensive programs of tax reduction, either as the result of voter referenda or of legislative action. The result has been growing deficits that have led to sharp reductions in state and local spending, even if these reductions cannot be sustained. Many states, in the face of declining revenues and cuts in federal aid, are now debating how they can raise revenues to pay for necessary programs. Yet, little is understood about how different types of state taxes shape local economic growth, nor about the most efficient ways of raising revenues. Although economists typically endorse the general principle of user fees, they are rarely specific on the way that tax revenues can be raised most efficiently by state

and local governments operating within the confused fiscal layer cake of state-federal relations. More specific advice that recognized the administrative as well as the economic aspects of revenue collection would be valuable.

In summary, the expanding role of state governments in managing local economic growth has not been matched by any growth in our understanding of how state policies can be used most effectively to attain public policy objectives. Academic research is needed to assist state policymakers during this period of rapid transition. But to prove useful, the research will have to be conducted in a way that takes into account the legislative and institutional framework within which state policy is conducted.

# THE ECONOMICS OF MASS MIGRATION FROM POOR TO RICH COUNTRIES

## Leading Issues of Fact and Theory

By MICHAEL J. GREENWOOD\*

Since the imposition of entry quotas in the 1920's, U.S. immigration issues have been of little concern to economists. First binding quotas, and later the effects of the depression and World War II, resulted in sharply reduced immigration compared to levels of the late nineteenth and early twentieth centuries. When, during the 1950's, immigration again began to rise toward quota ceilings, population was growing rapidly from other sources, and thus immigration continued to contribute relatively little to U.S. population growth. Moreover, during this period, mortality among the aging stock of foreign born more than offset net immigration, with the consequence that the stock declined by 4.6 million between 1930 and 1970. What attention was directed at international migration during this half century was mainly on the part of economic historians, who focused on the period of unrestricted flows, and on the part of those interested in the brain drain and concerned with the flow of high-level manpower from poor to rich countries.

Renewed interest among economists has recently been kindled in U.S. immigration issues, perhaps in part due to immigration having again taken its place as an important source of population growth. Historically, the years of heaviest U.S. immigration were early in this century. Between the 1900 and 1910 censuses, the foreign-born population increased by 3.2 million, which accounted for 19.5 percent of incremental U.S. population. Between the 1970 and 1980 censuses, the foreign-born population grew by 4.3 mil-

lion, which was 18.6 percent of the increment in population and the largest intercensal increase in U.S. history. After 1850, besides the 1900-10 decade, only two other decades show greater percentage contributions of the foreign born to population growth, 1850-60 (23.0 percent) and 1880-90 (20.6 percent). If we make a conservative assumption that illegal migration during the 1970's contributed 2.0 million additional persons to population who were unenumerated in the 1980 census, immigration accounted for 25 percent of the increment in U.S. population. Due to the increased importance of immigration, especially that from poor countries, a number of issues of fact and theory have become the focus of recent attention among economists.

### I. Issues of Fact

Facts concerning international migration are among the most elusive with which a social scientist must work. Differences in the definition and measurement of emigration and immigration make the formation of even a 2×2 matrix of international migration flows, for virtually any two nations in the world, doubtful as to accuracy and comparability. These problems caused me to avoid aggregating migrants from a group of countries categorized as poor to a group categorized as rich. Rather, I focused specifically on the United States as a destination country and examined official historical data on the source countries of immigrants. Because the data are official, they do not include estimates of illegal immigrants, who probably come predominantly from poor countries and whose numbers have probably risen dramatically over the past twenty years.

In considering the facts regarding migration from poor vs. rich countries to the

\*University of Colorado-Boulder. I am grateful to P. Lynn Stuart for carefully compiling the data on international migration that underlie the first part of this paper, and to Jane H. Lillydahl and John M. McDowell for helpful comments.

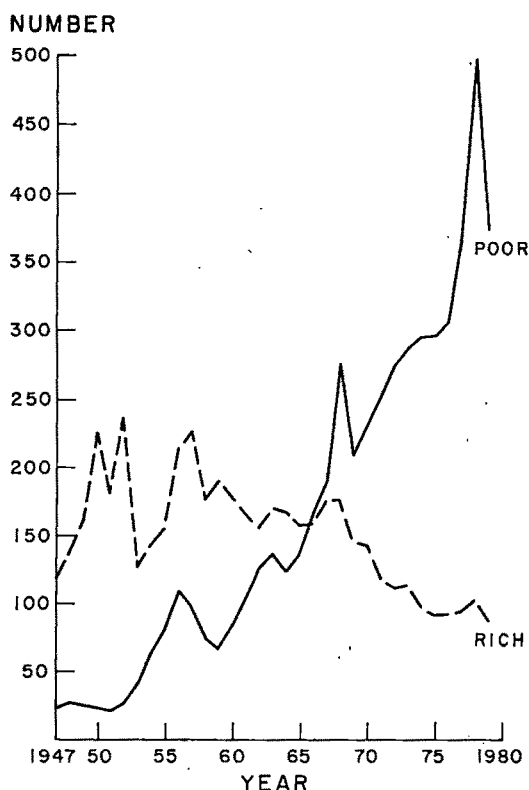


FIGURE 1. GROSS IN-MIGRATION TO THE UNITED STATES FROM POOR AND RICH COUNTRIES (IN THOUSANDS)

United States, I focused on two issues, namely, the source-country composition of the flows and the skill composition of the flows from each source country, where those who claim to be professional, technical, and kindred workers as well as managers and administrators (except farm) are categorized as skilled (and hence presumably rich compared to others in source countries). The dichotomy between rich and poor has been made as follows. Countries were ranked according to 1975 per capita income. For all countries listed in the National Accounts section of the United Nations *Statistical Yearbook, 1978/79*, median per capita income was approximately \$1,500, which was the figure chosen to separate rich from poor. Rich consists of Europe, Israel, Japan, Oceania, Canada, and Venezuela, whereas poor includes Asia (less Japan), Israel, Africa,

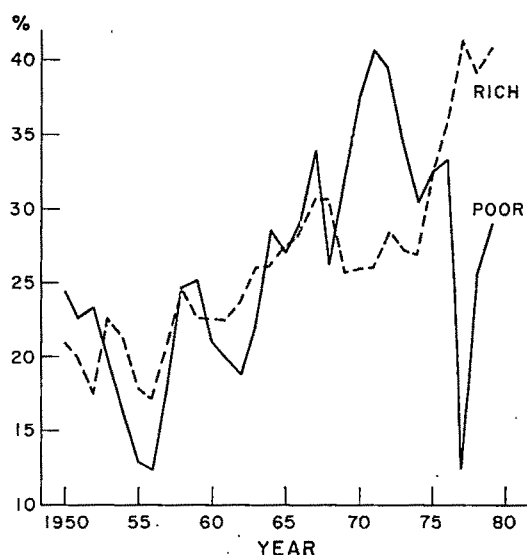


FIGURE 2. PERCENTAGE OF U.S. IMMIGRANTS CLASSIFIED AS SKILLED

North America (less Canada), and South America (less Venezuela).

Figure 1 shows that the gross influx of migrants from poor countries to the United States has grown steadily since 1951, with three periods of particularly heavy movement caused by flows from Mexico (1954-56), from Cuba (1966-68), and later from Cuba, Mexico, and Viet Nam (1977-78). This figure also shows that migration to the United States from rich countries, though erratic prior to 1960, has declined since then. Between 1949 and 1951, the annual average percentage of U.S. immigration originating in poor countries was 11.2, compared to a corresponding percentage of 81.3 for the period between 1977 and 1979. During the earlier period, migration from poor countries averaged 23,863 per year; compared to 413,202 per year during the later period. The Immigration and Nationality Act Amendments of 1965 had much to do with these changes.

As shown in Figure 2, the percentage of skilled immigrants displays marked fluctuations. Nevertheless, for rich countries the general trend is distinctly upward throughout the period, whereas for poor countries the trend is upward until 1971, after which it is

downward. The percent skilled from poor countries peaked in 1971 at 41 percent, then fell to 13 percent in 1977 when heavy movements from Cuba and Mexico occurred.

The shifting source-country composition of U.S. immigration has historical precedent. As the nineteenth century proceeded, the source of U.S.-bound immigration gradually shifted from northern and western to southern and eastern Europe. Joseph Spengler (1956) clearly feels that the earlier immigrants, who were largely from Great Britain and Germany, carried critically needed skills and know-how, whereas the later wave satisfied the large and growing demand for unskilled and semiskilled labor. Hence, earlier immigrants served as catalysts in U.S. development, whereas later immigrants played a more passive role.

Note a parallel shift in emphasis regarding more recent migration. During the late 1960's and early 1970's, when the drain of talent from poor to rich countries was a topic of interest, a prominent hypothesis was that poor countries, after investing in education and training of their citizens at considerable sacrifice, were losing those with critical skills to rich countries, where these individuals could earn higher returns. At the time, the general feeling was that a high proportion of international migrants were professional. This situation appears to have changed and, consequently, we seemingly find more emphasis being placed on migration of unskilled labor from poor to rich countries.

The data described above are characterized by two shortcomings that prevent them from fully reflecting certain of the most relevant facts concerning U.S. immigration. First, they show only that portion of the migrants who were legally admitted. They fail completely to reflect illegal immigration. Second, they relate only to gross migration, not to net migration. Since direct measures of neither the illegal component of U.S. immigration nor the emigration component of net migration are available, these components must be derived indirectly via demographic techniques. The resulting estimates have been of great interest and, dependent as they are on a number of underlying assumptions, have also been the subject of serious debate.

No fact concerning international migration has been more prominent in recent years than that concerning illegal migration to the United States. This entire issue has two main topics of concern. The first has to do with the stock and/or flow of illegal residents, whereas the second has to do with the permanency of residence. Note that the question of permanency has direct bearing on the question of emigration, and thus has immediate relevance for net migration. A number of attempts have been made to use scientifically defensible methodologies to estimate either the stock of illegal residents in the United States, or their net flow. Mexico has been the primary focus of these efforts, due primarily to the fact that in recent years almost 90 percent of the deportable aliens apprehended by the Border Patrol have been of Mexican origin. The number of deportable aliens apprehended yearly averaged approximately 673,000 during the 1970's, and was well over one million in each of the last three years of the decade. The Census Bureau estimates that between 3.5 and 6.0 million illegal aliens reside in the United States. Given the clandestine nature of illegal migration, the fact that estimates of its magnitude are even remotely similar may be surprising. This entire issue, however, is one that is unlikely to be definitively resolved.

Guillermina Jasso and Mark Rosenzweig (1982) consider the issue of net migration. They do so by deriving estimates of net U.S. emigration rates for the 1971 cohort of legal U.S. immigrants, by country of origin. They develop upper- and lower-bound estimates of rates of net emigration over a period of eight years after immigration. Their main findings are that emigration rates of prior immigrants are substantial, perhaps as high as 50 percent, and vary systematically according to the migrants' original source country, with particularly high emigration rates for Canada, Central and South America, and the Caribbean (less Cuba), but low rates for Asian countries. Although the ranges between their lower- and upper-bound estimates are substantial (for example, 15.6 to 56.2 percent for Mexico), the finding that proximity (and hence cost of remigrating), as well as political and economic climate, are important de-

terminants of emigration after prior immigration is of great interest and lends credence to the results. The range of estimates identified by Jasso and Rosenzweig also emphasizes the inherent difficulty in developing a definitive estimate of net migration, which, like its gross illegal component, will likely remain a highly debatable issue.

## II. Issues of Theory

At least as reflected in the literature, issues dealing with pure theory are not currently among the leading issues concerning international migration. Theory is far ahead of verification, and the major issues of the time concern verification. Hence, in what follows I interpret "theory" broadly.

Among the most controversial issues regarding the impacts of immigrants on the receiving country are those concerning employment opportunities, wages, and working conditions of domestic workers. Do immigrant workers displace domestic workers from jobs, cause a reduction of domestic wage rates, and encourage undesirable working conditions? These questions are difficult to resolve because immigration may have negative effects on certain workers, while it has positive effects on others.

The idea underlying the debate is that the level of employment ( $E$ ) at the end of a given period  $t$  is dependent upon the amount of net employment migration ( $M$ ) that occurs during  $t$ . Thus,  $E_t = f(M_t, X_t)$ , where  $X_t$  is a vector of variables other than migration. The partial derivative  $\partial E_t / \partial M_t$  shows the employment effect of one additional employed migrant. This value (hereafter  $= \beta$ ) can be greater than, equal to, or less than one, depending upon whether one more employed migrant causes employment to increase by more than one job, by one job (the segmentation hypothesis), or by less than one job (the replacement hypothesis).

Under free-market conditions, migration appears frequently to be a self-reinforcing and cumulative phenomenon because it increases labor demand as well as labor supply in receiving areas. (Fragmentary evidence on internal U.S. migration suggests that  $\beta \geq 1$ .) Many factors may underlie such a relation-

ship. Among these are 1) the skills, inventiveness, and innovativeness of the migrants themselves, who may possess differential endowments of human capital relative to the population of the sending or receiving area. Spengler appears to have had this situation in mind when he discusses the contributions made by northern and western European peoples to U.S. development. Although others presently hold a similar position, this alternative seems less applicable when immigrants are mainly poor, unskilled individuals.

Apart from their human capital, 2) migrants may own physical and financial capital that they bring with them. 3) Migrants may possess sources of income other than their labor services. 4) Migrants may cause investment in receiving localities. 5) Migrants may influence the price of locally produced goods and services due to the changed demand they may cause for such goods and services. 6) Migrants may contribute to the growth of markets and to the achievement of scale and agglomeration economies. Many economic historians believe that this last factor was operable in the United States up to about 1900 or 1920, but has not been important since then. Each of the above factors thus appears to cause a dampened employment effect from a poor relative to a rich immigrant.

Regardless of the value of  $\beta$ , the distributional consequences of immigration have been a serious issue. Even if  $\beta \geq 1$ , low-income, poorly educated, unskilled immigrants may be competing primarily with comparably educated and skilled U.S. residents. One view is that the jobs filled by immigrants from poor countries are of no interest to U.S. citizens, but little concrete evidence is presented to support this position. The other view is that low-income U.S. residents, primarily blacks, Mexican Americans, and Puerto Ricans, are hurt by job competition from immigrants. Historical precedent exists for this latter position. After the Civil War, the flow of immigrants to northern U.S. cities appears to have fluctuated inversely with the flow of blacks out of the South and into these cities, as well as with the flow out of rural areas in general.

The effects of immigration also appear to be dependent upon the timing of the migration over the cycle, with the negative effects on poor domestic workers being most pronounced during periods of high unemployment. Finally, much attention has been directed at specific U.S. regions that have been particularly impacted by immigrants.

In general, issues dealing with economic growth have received the most emphasis in North America. Immigration's effects on inflation and the balance of payments have not been at the center of the debate. These latter two topics are much more widely discussed with reference to European migration. Moreover, except for research on the earnings behavior of immigrants and their offspring, little emphasis has been placed on the short-run vs. long-run benefits and costs of North American immigration.

Has international migration helped or hindered the economic development of poor source countries? Economists have generally taken the position that migration is beneficial to countries of emigration. The brain-drain debate, however, focused on the negative consequences of outmigration of the most highly educated. These consequences were identified in terms of the opportunity cost of investments in education and training, or of the lost talent made possible by such investments. Recently the emphasis has shifted, probably in response to migration from poor countries having become more heavily oriented toward migration of poor people than was true in the late 1960's and early 1970's.

One important issue on which little agreement exists deals with the effects of remitted earnings on countries of emigration. Though their precise magnitude is difficult to measure, remittances appear frequently to be substantial, and they also appear to be closely

related to the permanency of the move, falling in magnitude and regularity with increased duration of the migration. In this sense, remittances are closely related to the conditions that give rise to the migration. Whether the remittances are channeled into investment or are used primarily for current consumption seems to depend upon economic and social conditions in the countries of emigration.

A second important issue deals with the effects of skills learned abroad by migrants who later return to a poor country. Little is known about the applicability of skills learned in a rich country to a poor country of origin. An hypothesis central to the emerging literature on this issue is that a labor force oriented toward agricultural skills can be transformed into one oriented toward urban industrial skills through a period of employment in an industrialized economy. If Jasso and Rosenzweig are correct in concluding that Western Hemisphere migration to the United States is more likely to entail later emigration, the issues of remitted earnings and remitted skills will assume an importance in the Western Hemisphere similar to that accorded them regarding international movements elsewhere in the world.

## REFERENCES

- Jasso, Guillermina and Rosenzweig, Mark R., "Estimating the Emigration Rates of Legal Immigrants Using Administrative and Survey Data: The 1971 Cohort of Immigrants to the United States," *Demography*, August 1982, 19, 279-90.
- Spengler, Joseph J., "Some Economic Aspects of Immigration into the United States," *Law and Contemporary Problems*, Spring 1956, 21, 236-55.

# International Migration Models and Policies

By EDWIN P. REUBENS\*

The problems of international mass migration, which are vehemently debated among the directly interested parties and in the U.S. Congress, have been curiously neglected in the economic literature, in most of the American Economic Association meetings, and even in the United Nations current demands for a New International Economic Order. My own recent book (1981) is almost alone in devoting a chapter to migration problems that are important in North-South relations.

Most economists seem to have accepted the bias of the orthodox international theorists, namely that goods—and even services, capital, and technology—move across national borders in accordance with comparative advantage, while people are immobile. Yet this bias is in simple defiance of the facts. Millions of persons annually migrated in the years before the Great Depression; then World War II almost halted the movements. More recently, the numbers have soared again, boosted by “guestworker” practices in Europe during the 1960’s, and by floods of refugees into the United States in the late 1970’s, and by hundreds of thousands of “undocumented” migrants, primarily workers coming from Mexico across the U.S. southern border. These inflows, which currently add annually only 1 percent or less to the population and labor force of most of the receiving countries, are in fact minor magnitudes compared to the potential for multimillion flows from the poor countries of the world; flows which would vastly intensify the effects upon labor supply, unemployment, wages, social stresses, economic development, international comparative advantage, balances of international payments, and almost every feature of economic and social life.

While the “brain drain” attracted some welfare theorizing in the 1960’s and early

1970’s, in more recent years a few economists—some of revisionist or radical persuasion, others internationalists, others econometricians—have begun to analyze the mass migration phenomena and policy problems. But their arguments still tend to rely on tautologous formulations and simplifying assumptions drawn from traditional mechanistic market theory, in which people are treated as mindless particles pushed and pulled by irresistible forces. Sadly lacking in support are empirical studies of migrants in actual market structures and behavior patterns (here the research program of the recent Select Commission on Immigration and Refugee Policy was a great disappointment). Meanwhile the several parties vitally interested in migration have been hard at work to set or change actual policies—seeking to open or close the immigration doors—and putting to their own uses various bits of economic theory and doctrine, many of which had been designed for different contexts, even contrary purposes, as shown below.

In view of these ambivalences and paradoxes in the current arguments on migration, I will first pick up the diverse aims or policies, since they are usually predetermined, and so proceed to the respective models involved in those policies. This approach is not to denounce models—which are essential, of course, for choosing variables and interrelating them—but rather is to analyze the current doctrines for internal logic and consistency, point out their biases, and test for realism and completeness. Five distinct doctrines are noted, and designated according to their access features, or “doors” for emigration and immigration, along with the respective underlying models.

## I. The Open Door and the Classical Model

The first doctrine, here termed the “Open Door” or “Open Borders” policy, used to be called “internationalist” (see H. G. Grubel

\*Professor of economics, City College of New York.

and A. D. Scott, 1966, and Walter Adams, 1968). Originally concerned with small numbers of migrant professionals, this doctrine is expanded nowadays to include the bold new conception of "economic refugees," who may number in the millions. Internationalist doctrine envisages virtually unlimited, universal admission of all applicants, in whatever types and numbers. It rejects or ignores the national migration restraints that everybody else takes for granted. It draws, on the one hand, upon humanitarian, egalitarian, and civil liberties principles; on the other hand, it derives from an absolute commitment to the operation of free markets for optimizing the allocation of total, worldwide resources.

In the real world, however, with vast economic disparities between the more and the less developed countries (*MDCs* and *LDCs*), the intended unlimited flows would make indeed a drastic jump to worldwide equalization. It would probably be a zero sum, or even a negative sum transaction, as employment and consumption levels in the *MDCs* tumbled violently, to make a small individual accommodation for the millions of poor migrants from the *LDCs*. Long before this came about, the *MDC* citizens would react to slam shut that open door.

In fact, that model quickly runs into the ultimate lesson of classical economics, which envisages worsening scarcity, growing distributional stresses as real wages fall and rents rise, and the eventual end of capital formation. This lesson, which Ricardo and Malthus drew from the natural increase of local population, may here be extended to the impact of rising international migration under any regime of open borders and high mobility amid limited growth in aggregate employment and income.

## II. The Closed Door and Nationalistic Neoclassicism

The term "Closed Door" refers to a zero quota imposed on migrational exits or admissions among nations. The economic model here is usually a nationalistic version of the neoclassical economy, closing off the international market in labor while extolling the virtues of the market within each impregna-

ble country (pursuing a *national* instead of a *world* social welfare function).

One motive for such a policy is sheer xenophobia, sometimes dressed up in terms of "preserving cultural purity" or "safeguarding our way of life." Another aim is to limit population pressures on "the environment." A more economic motive is to prevent competition by "cheap foreign labor" here, in employment, wages, working conditions, and use of social services. All of these are defensive aims, adopted by "risk avoiders" that are therefore automatically opposed to immigrants, who are essentially "risk takers."

Sincere neoclassicists who find themselves putting up barriers against foreign workers are troubled by their own nationalistic retreat from economic competition and general liberalism, and seek various justifications. One such justification takes shelter in Pareto optimality, permitting changes only if they will not reduce the welfare of any persons anywhere. Such welfare economists—desperately trying to avoid interpersonal comparisons—end by closing their eyes and ears to all the important problems of our time, or rejecting attractive proposals simply because they may impose some loss somewhere (often on some group that is already advantaged!).

Perhaps the most interesting is an economic rationale along the lines of macro-micro synthesis. This is an assertion that immigration generally lowers real income per capita in the receiving country, because a given total income must be divided among more claimants.

That line of argument depends crucially upon the neoclassicists' simple and unquestioned assumptions, namely that full employment of productive resources prevails in the receiving country, wages being flexible, and there are no persistent job vacancies, whether overall or in particular locations and segments (notable in low-level job segments), and there are no prospective technological advances nor even economies of increasing scale, nor a return to rapid growth. Given such assumed circumstances, foreign entrants must compete with and often displace native workers, or at least reduce their wages while raising the profits of employers.

But these are static and artificial conditions, amounting to a fixed "jobs fund," and implying a predetermined "optimal population" which should not be exceeded. If these conditions are seen as unrealistic, or at least as alterable, most of the nationalistic neo-classical argument for simply excluding immigrants will collapse. (Barry Chiswick, for example, permits even unskilled immigration to raise the aggregate income of the natives in his production function.)

### III. The Screen Door and Pressure-Group Economics

Most of the *MDCs* maintain an immigration system in the moderate range between the open and closed door policies, namely a national screening system that admits limited types and number of foreigners; using various devices such as the "quota-and-preference" regulations of the United States, or the "points" system used in Canada, or the "guestworker" programs in Western Europe (comparable to the H-2 and Exchange Visitor U.S. programs).

The underlying rationale is seldom stated, and even more rarely examined. Instead, a loose vocabulary of "push and pull" factors commonly obscures the issues. But what we see in fact are administrative systems for which no specific maximands have been set. They do incorporate—implicitly if not explicitly—several different aims of reuniting family members, and of meeting particular labor shortages, and of providing political asylum. While they impose some constraints, little or nothing is said in immigration circles about the economists' objective functions of promoting capital formation, or of maximizing output in total or per capita, stabilizing the price level, or equilibrating international payments. Only the aim of minimizing unemployment has recently become important.

As for the aims actually pursued, immigration legislators manage to avoid specifying the respective magnitudes sought, or the importance of these aims relative to one another, or the prices we are willing to pay to implement or reconcile these objectives.

In consequence, our screen door immigration systems are a hodge podge mostly

brought about by various special interest groups: notable employers and labor unions (sometimes opposing each other on immigration, sometimes agreeing on "sweetheart" deals), as well as organized ethnic associations (which also have pro- and anti-immigration subgroups), also the friends of various political refugee groups. In economic terms, immigration is determined by strategic games played by small pressure groups, rather than consensually decided by political or market interactions of our whole electorate or labor force (compare Mancur Olson, 1982).

Usually, however, the screen door swings arbitrarily, and rattling worst when complex structural or dynamic matters are subjected to simplistic and static neoclassical principles and mechanisms, or vice versa. Some labor leaders (for example, Sol Chaikin, I.L.G.W.U., and Robert Harbrant, Food Dept., AFL-CIO) and their academic allies (for example, Ray Marshall; Vernon Briggs) contend that immigrant workers "undercut" the jobs, the wages, and the working conditions of American labor; and their policy is to demand reductions in immigration, and prohibitions on the employment of illegals. But, in fact, these contentions misconstrue the still small relative numbers of immigrant workers currently, and their weak impact on the aggregate labor market, their useful slowdown effect on declining industries, and the limited enforceability of prohibitions. In the particular segmented markets where immigrants concentrate—chiefly in low-level work—the unionist opponents of immigration usually ignore the phenomena of non-competing groups. Finally, those who invoke the record-high unemployment of Americans in the recession of 1981–82 are in effect demanding long-term and structural prohibitions when short-term adjustments may be all that are required.

A related example of modelling confusion occurs in parts of the theory of a "secondary labor market," in which both native minorities and foreign workers are said to be "confined," and therefore are said to be in fierce and destructive competition with each other. But these theorists, whose policy is to demand liberalization of the labor market, and

upgrading of low-level jobs, neglect the economic forces tending to perpetuate those jobs at low pay, or to curtail them if pushed to higher pay. Also neglected is the ability of most native workers to avoid the low-level jobs so long as they can draw upon other sources of support.

In a third area, some spokesmen for the *LDCs* make the neoclassical claim that emigration removes scarce workers (both skilled and others, amounting to a "brain-and-brawn drain"), and thereby reduces output and raises wages, so that economic development, welfare, and competitive strength are impaired. They demand either prevention of that "drain," or compensatory payments to the *LDCs*. However, they are treating those *LDCs* as if labor were scarce there, rather than actually redundant, with social marginal productivity approaching zero—as observed on both the professional and the unskilled levels. Likewise ignored are the important remittances which emigrants send home in foreign exchange.

All these arguments, which adopt some realistic structural and behavioral features, but apply idealized market mechanisms to deduce policy conclusions, exemplify the classic philosophical confusion of Aristotelian particularist substances with Platonic abstract ideas.

#### IV. The Retaining Door and Developmental Models

Virtually all the aims at issue here, whether for or against immigration, converge on policies to promote economic development in the *LDCs*, the chief source of persons migrating to escape from poverty and underemployment. While the models vary greatly as to the prime "engine of growth," what they have in common is the belief that economic development will soon absorb and retain all of the huge redundant labor supply, and so preclude emigration.

However, economic development is a very slow process, in which output seldom grows at more than 5 percent a year over the long term, nor at more than 8 percent in favorable short periods. Furthermore, modern development tends to be highly capital intensive,

such that the expansion of employment is at a still slower rate than expansion of output. Often more jobs are destroyed. While man-hours of productive employment may expand at 2–3 percent a year, population is often growing at least as fast, and the labor force even faster, thereby tending to maintain if not markedly increase the surplus of labor in such countries. Note the case of Mexico for an extreme example of huge and spreading underemployment amid rapid growth of *GNP*.

In the very long run, development may dry up the labor surplus and raise wages in tropical agriculture, and so improve the terms of trade for *LDCs*, as Arthur Lewis has suggested (1978). But for the decades to come, despite even the best efforts for development, the disequilibrium system will persist: millions of persons in Asia, Africa, and South and Central America will still be seeking to migrate, and will be aided by the very rise of incomes and aspirations that accompany development.

#### V. "No Doors—Demolish the House": The Imperialism Model

I come now to the Marxist/Leninist critique of most of the foregoing doctrines on migration. Viewing those as mostly apologia for class exploitation at home and economic imperialism abroad, the Marxist/Leninists charge that the "capitalist class" imports or "recruits" foreign workers to assure cheap labor, and to undermine labor organization and revolutionary action. At the same time, those critics view the exportation of capital and enterprise to the *LDCs* (especially via the "multinational firms") as a means of taking unfair advantage of cheap labor there (making cheap products that can largely be shipped home), while displacing in those *LDCs* many traditional workers who become available for emigration (a neatly closed vicious circle!).

Other features recently added to this basic model include "increasing commodification of labor," domination by the "center" nations over the economy and politics of the "periphery" *LDCs*, and increasing economic instability in the capitalist nations, leading to

the expansion of the "secondary sector" and enlarged requirements and recruitment for imported labor.

These radical models—ideological, Procrustean, and scornful of divergent facts—have become an article of faith and self-assurance for innumerable writers on migration, especially in Latin America. For these writers, who are mostly noneconomists, the radical models exert a great appeal by simplifying complex problems into simple macro categories, and by offering striking images on the "seesaw" principle or zero sum accounting, and they dramatically present a nominal villain—the bourgeois ruling class—to hate and destroy.

When it comes to the evidence for testing these radical models of migration, these doctrines are sadly deficient. The alleged evidence is often merely anecdotal, frequently distorted, stressing damages while neglecting benefits, and at worst is quite solipsistic (endlessly repeating their beliefs but not testing them). The authentic points they sometimes present have mostly been incorporated into more valid models of development, intersectoral shifts, and the gains from trade.

## VI. Towards a Better Model of Migration

Noting the deficiencies of many current doctrines on emigration and immigration, I see a need for integrating various elements from the foregoing models into a new two-fold paradigm: a *casual* model of Aspirations/Opportunities/Mobility, in which the migrant is viewed as an entrepreneur disposing of his human capital as best he can; and a *consequences* model of Needs-for-foreigners and Capacity-to-absorb-them, in which immigration is treated by analogy to foreign capital inflows. In the *LDCs*, the chief variables would be rising personal aspirations, persistent welfare disparities with the *MDCs*, and growing economic ability to afford migration. In the *MDCs*, the chief variables would be labor shortages for low-level jobs, and idle capacity of some social facilities in some locations.

A reconsideration of the Simpson/Mazoli immigration reform bill, now before Congress, in the light of such a paradigm, would make it clear that this bill is mainly administrative reform, with its amnesty for long-resident illegals and its prohibitions on employing others. What it omits or understates are the more economic solutions, namely: (a) provisions for comprehensive measurement of U.S. labor market needs, that would take full account of low-level jobs and American workers' preferences; (b) a program of internal retraining and relocation of American workers, geared to local differences in both job vacancies and the capacity to absorb migrants; (c) an adequate program for temporary foreign workers in variable magnitudes adjusting to the changing U.S. labor market, and so precluding illegal inflows by legitimately filling up most of the vacancies that attract them; and (d) a multinational negotiating system to adapt the present unilateral rules on emigration/immigration for the changing conditions in the less developed and more developed countries.

## REFERENCES

- Adams, Walter, *The Brain Drain*, New York: Macmillan, 1968.
- Chiswick, Barry R., *The Gateway*, Washington: American Enterprise Institute, 1982.
- Grubel, H. G. and Scott, A. D., "The International Flow of Human Capital," *American Economic Review Proceedings*, May 1966, 56, 268-74.
- Lewis, W. A., *The Evolution of the International Economic Order*, Princeton: Princeton University Press, 1978, ch. 3.
- Olson, Mancur, *Economic Growth, Stagflation, and Social Rigidities*, New Haven: Yale University Press, 1982.
- Reubens, E. P., "International Migration in North-South Relations," in his *The Challenge of the New International Economic Order*, Boulder: Westview Press, 1981.

# Trade Theory, Distribution of Income, and Immigration

By FRANCISCO L. RIVERA-BATIZ\*

The mass migration of unskilled labor from poor to rich countries has become a subject of intense concern in the rich countries. This can be noticed from the growing public outcry on the topic, and from the vocal debates on immigration policy thriving among policymakers. This paper examines the impact of mass immigration utilizing the analytical framework of so-called trade theory. As such, its emphasis is on general equilibrium effects and on the constraints imposed by the openness of the economy on such effects. The analysis will center on the distributional impact of immigration. This is a matter which has generated much controversy, as some groups in the economy (such as labor) have opposed immigration on the basis they are hurt by immigrants, while others (such as producers facing intense import competition from developing countries) have lobbied for increased immigration on the basis they may not subsist without them.

## I. Output Composition Effects and Immigration

In spite of the popular belief that immigration turns income distribution against labor and in favor of capital, standard trade theory offers the strong presumption that immigration may actually have no significant effects on income distribution at all. Consider the standard trade model of a small economy producing two traded goods,  $X$  and  $Y$ , under conditions of perfect competition and constant returns to scale. There are fixed endowments,  $\bar{K}_n$  and  $\bar{L}_n$ , of nationally owned capital and labor, which are perfectly mobile among sectors in the economy and operate in perfectly competitive factor markets. These define the national transformation curve  $AZ$ , shown in Figure 1. With an international

price ratio given by  $\bar{P}\bar{P}$  (whose slope is  $-P_X/P_Y$ , with  $P_X$  and  $P_Y$  being the prices of  $X$  and  $Y$  in international markets), the production point of the economy would be at  $E$ . The "income distribution locus" for labor is given by  $WM$ . This curve shows the share of output that national labor would receive (earn) at different points on the national transformation curve. For instance, given production point  $E$ , we can draw the ray  $OE$  from the origin to that point. The  $WM$  locus is constructed so that the intersection of the ray  $OE$  with  $WM$  (at point  $C$ ) gives us the relative claim of labor on national output. The length of the segment  $OC$  represents labor's claim on output, while  $CE$  represents capital's claim;  $OC/CE$  then represents the factor share of labor relative to capital. Similarly, other points along  $WM$  represent labor's factor share associated with other production points along the  $AZ$  curve. The shape of  $WM$  is determined by the behavior of factor shares as output is varied along  $AZ$ . The factor earnings of capital relative to those of labor are given by  $\theta = r\bar{K}_n/w\bar{L}_n$ , where  $r$  and  $w$  are the factor prices of capital and labor. Since national factor endowments are fixed,  $\theta$  will change only in response to changes in relative factor prices. As drawn, labor's share relative to capital's decreases as production of  $X$  increases. The explanation lies on the assumption that sector  $X$  is relatively capital intensive. As output of  $X$  increases, this sector's demand for labor will not increase by as much as sector  $Y$  releases its own labor. Excess labor supply results and the wage-rental ratio decreases.

What is the impact of immigration in this framework? An increase in the aggregate labor endowment would shift the economy's transformation curve upwards, as illustrated by the move from  $AZ$  to  $A'Z'$ . Aggregate production would then move to  $E'$ , where  $\bar{P}\bar{P}$  is tangent to  $A'Z'$ . Note, however, that the equilibrium of the national residents, lying along the national transformation curve

\*Department of economics, University of Chicago. I am indebted to Luis Rivera-Batiz for useful comments. The research embodied in this paper was supported in part by a grant from the National Research Council of the National Academy of Sciences.



price disturbance. The net balance of these two effects on individuals depends on their factor endowments and tastes. If the economy is composed of a group of rich capitalists who consume mostly commodity  $Y$ , and a group of poor workers who consume mostly  $X$ , then the size distribution of income will clearly worsen. Note finally that if immigration is accompanied by a *decrease* in the relative price of  $X$  in terms of  $Y$ , then the distributional impact will be to benefit labor and hurt capital!

Relative prices can be affected even if the economy faces fixed terms of trade. This occurs because of the presence of (internationally) nontraded goods, such as services, whose prices are determined by local demand and supply. Actually, the impact of immigration on the nontraded goods sector may be of immense importance, due both to the size of the sector in most developed countries (composing as much as 56 percent of *GNP* in the United States), and because of its significance in employing immigrants (more than one-third of all legal immigrants entering the United States in 1977, for example, declared occupations associated with nontraded goods sectors).

Assuming fixed terms of trade, we can aggregate exportables and importables into a composite traded good which can be represented by  $X$ . If  $Y$  represents nontraded goods, Figure 1 can still be used as a tool of analysis. Suppose point  $E$  is the premigration equilibrium and that, with nontraded goods prices unchanged, immigration moves aggregate production to point  $E'$ . This would reduce the economy's output of traded goods ( $X$ ) and increase that of nontraded goods. The additional supply of nontraded goods would then induce a reduction in their price, which is shown by the switch from  $\bar{P}\bar{P}$  to  $\bar{P}\bar{P}'$ . The final production point for the nationals would be at point  $J$ . As a consequence, income distribution is affected by the same considerations noted earlier in this section. Functional income distribution, for instance, would turn against labor and in favor of capital. In addition, those individuals who consume relatively more nontraded goods would tend to benefit (see my 1982b article for details; these results must be qualified

when there are more than two industries and factors, see my 1982a paper).

### III. Sector-Specific Factors and Income Distribution

It was shown above that output composition shifts can prevent immigration from affecting income distribution. The operational importance of this effect, however, depends on how easily inputs can be shifted among sectors in the economy. Within short periods of time, it may be reasonable to expect capital to be relatively immobile. If capital is considered sector specific, then, in effect, there will be three inputs in the economy:  $X$  industry capital,  $Y$  industry capital, and labor, and three factor prices to analyze: the rental rates in each sector, and the wage rate. In this context, immigration will tend to decrease the wage rate and increase rental rates in each sector. This is because the sector-specific capital generates diminishing returns at the economy-wide level (Michael Mussa, 1974). The additional workers decrease the marginal productivity of labor, and increase that of capital in each sector. Consequently, the distribution of earnings will turn against labor and in favor of capital. The size distribution of income will also be affected, with those individuals owning more capital relative to labor gaining, not only from the discussed shift in factor prices, but also because capital owners receive a (scarcity) rent on the use of the immigrant labor.

When sector-specificities characterize only the short run, capital will move over time into production of the labor-intensive commodity. Wages will then increase and rental rates decrease back towards their original levels. Long-run equilibrium will involve no distributional effects; only over the adjustment process will income distribution change (see my 1982c article, and A. Sapir, 1983).

Up to this point, I have considered labor to be perfectly homogeneous and freely mobile. This contrasts with many applied studies on immigration, which tend to view the labor market in the recipient countries as being segmented into quite distinct sectors, each characterized by a quite distinct type of labor (Michael Piore, 1979). Accordingly, let

us suppose that sector  $X$  uses only skilled labor while sector  $Y$  uses only unskilled labor. If mass immigration of unskilled workers occurs, the wage rates of both skilled and unskilled labor decrease and the rental rate on capital increases. As in the case of sector-specific capital, the economy is subject to aggregate diminishing returns, implying the additional unskilled workers must reduce their wage rate and increase capital's rental rate. The wages of skilled workers, on the other hand, decrease because the initial impact of the immigrants is to increase the rental rate in sector  $Y$ , inducing capital to leave sector  $X$ . With less capital to work with, skilled workers will face lower marginal products, and thus lower wages. So, even if nationals are predominantly employed in industries where there are few immigrants, they will still not necessarily be insulated from the effects of immigration.

In contrast to the many restrictions governments impose on immigration, international capital mobility seems to be relatively more free from controls. If capital is freely mobile among countries, what difference does this make in terms of the present discussion? As it has been shown, unskilled labor immigration increases the rental rate on capital. This would then induce an inflow of foreign capital into the economy, which would partially reverse the losses of labor relative to capital.

#### IV. Immigration in the Presence of Distortions

Most economies are plagued by economic distortions, originating from both private and public manipulation of the marketplace. Examining the distributional impact of immigration under these conditions is thus imperative. I consider first immigration in the presence of a tariff, assumed to be placed on the (labor-intensive) importables. In terms of Figure 1, the tariff-inclusive price is now represented by means of  $\bar{P}\bar{P}$ . Consequently, the economy's initial production point is  $E$ . Immigration would then move the economy's equilibrium to  $E'$ , but leave national production at point  $E$ . Factoral income distribution would remain unchanged. The size distribution, however, will tend to worsen. Im-

migration increases the economy's output of importables, and thus tends to reduce imports. This reduces the tariff revenues collected by national customs authorities. As a result, the nonimmigrants, to whom it can be assumed the revenues are distributed, will face a reduction of their real income. Given that domestic output of importables is still produced at the artificially high price imposed by the tariff, the loss in the real income of nationals must be collected by someone else. It is actually received by the immigrants, who are paid at a marginal product whose value is inflated by the tariff-inclusive price, and is thus higher than its worth at international prices (see Bhagwati, 1982). Given the loss of tariff revenues, and if the allocation of these is made on the basis of income, then those individuals at the lower tail of the income distribution, who benefit relatively more from the revenues, will lose relative to high-income groups.

The presence of factor market distortions brings into focus the analysis of unemployment. A useful scheme to describe some aspects of labor markets in developed countries is to decompose the economy into a unionized sector that sets artificially high wages and generates unemployment, and a competitive sector with flexible wages and full employment. Labor market equilibrium is assumed to occur when the expected wage in the protected sector (equal to the union wage weighted by the probability of employment) is equal to the wage in the competitive sector. Immigration of labor into the competitive sector lowers its wage rate, inducing nationals to migrate to the unionized sector until the increase in unemployment in that sector lowers the expected wage rate enough to maintain parity with the competitive wage. As a result, nonunion workers will lose relative to union workers. This loss occurs in the form of lower wage rates in the competitive sector and higher unemployment rates in the unionized sector (see my 1981 article).

#### V. Conclusions

Some crucial aspects regarding the distributional impact of unskilled labor immigration have been examined. The weight of the

analysis is that immigration does tend to turn the functional distribution of income against labor and in favor of nonlabor factors. This result, however, is not exactly obvious, and depends on the importance of output composition effects, international capital movements, and changes in relative prices, among other things.

## REFERENCES

- Bhagwati, Jagdish N., *Import Competition and Response*, Chicago: University of Chicago Press, 1982.
- \_\_\_\_\_, and Brecher, R., "Foreign Ownership and the Theory of Trade and Welfare," *Journal of Political Economy*, June 1981, 89, 497-511.
- Jones, R. W., and Scheinkman, J., "The Relevance of the Two-Sector Production Model in Trade Theory," *Journal of Political Economy*, October 1977, 85, 909-35.
- Kenen, P. B., "Migration, the Terms of Trade and Economic Welfare in the Source Country," in J. N. Bhagwati et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam: North-Holland, 1971, 238-60.
- Mussa, M., "Tariffs and the Distribution of Income," *Journal of Political Economy*, December 1974, 82, 1191-203.
- Piore, M., *Birds of Passage: Migrant Labor and Industrial Societies*, Cambridge: Cambridge University Press, 1979.
- Rivera-Batiz, F., "The Effects of Immigration in a Distorted Two-Sector Economy," *Economic Inquiry*, October 1981, 19, 626-39.
- \_\_\_\_\_, (1982a), "Nontraded Goods and the Pure Theory of International Trade with Equal Numbers of Goods and Factors," *International Economic Review*, June 1982, 23, 401-09.
- \_\_\_\_\_, (1982b), "International Migration, Nontraded Goods and Economic Welfare in the Source Country," *Journal of Development Economics*, September 1982, 11, 81-90.
- \_\_\_\_\_, (1982c), "On the Distributional and Welfare Impact of Immigration: Costs and Benefits," mimeo., University of Chicago, October 1982.
- \_\_\_\_\_, "The Economics of the 'To and Fro' Migrant: Some Welfare-Theoretic Considerations," *Scandinavian Journal of Economics*, 1983, forthcoming.
- Sapir, A., "Foreign Competition, Immigration and Structural Adjustment," *Journal of International Economics*, 1983, forthcoming.

## DEREGULATION, COMPETITION, AND EFFICIENCY

### Marginal vs. Average Cost Pricing in the Presence of a Public Monopoly

By DONALD J. BROWN AND GEOFFREY M. HEAL\*

The Arrow-Debreu analysis of decentralized resource allocation in a Walrasian economy assumes constant or decreasing returns to scale in production. Recently, several authors have extended this analysis to economies with a public monopoly, that is, a firm with increasing returns to scale. In this literature, the salient feature is the characterization of increasing returns to scale technologies as nonconvex production sets, so that under this definition both single and multi-product firms may exhibit increasing returns.

Here, our intended model is an economy with a competitive sector consisting of households and firms with convex technologies, and a public sector consisting of firms with nonconvex technologies. A special case is a single multiproduct firm which produces products for regulated markets (with a nonconvex technology) and produces products for unregulated markets (with a convex technology), for example, AT&T. We consider for such an economy two of the general equilibrium concepts that have been investigated in this literature. One is Harold Hotelling's notion of a marginal cost-pricing (*MCP*) equilibrium, and the other is M. Boiteux's notion of an average cost-pricing (*ACP*) equilibrium.

A marginal cost-pricing equilibrium is a family of consumption plans, production plans, prices, and lump sum taxes such that households are maximizing utility subject to after-tax income; firms with constant or de-

creasing returns are maximizing profits; the public monopoly is pricing at marginal cost, where potential losses are covered by the lump sum taxes; and all markets clear.

An average cost-pricing equilibrium is a family of consumption plans, production plans and prices such that households are maximizing utility subject to their budget constraint; firms with constant or decreasing returns are maximizing profits; the public monopoly is pricing at average cost, that is, breaking even or making zero profits; and all markets clear.

Unfortunately, all of the extant proofs of existence of a *MCP* or an *ACP* equilibrium are somewhat technical in nature and lack the transparency of counting equations and unknowns which many economists accept as an intuitive, if not formally correct, proof of existence. In view of this, one of the purposes of this paper is to demonstrate the existence of a *MCP* and an *ACP* equilibrium in a simple economy with increasing returns, where the equilibrium notions are characterized by systems of behavioral equations and market-clearing conditions. We give both an intuitive proof of existence by counting equations and unknowns, and a formal argument that these systems of equations have a solution by use of a simple fixed-point argument.

In addition, we review several of the standard partial equilibrium prescriptions for the regulation of a public monopoly and show that in a general equilibrium model they can be interpreted as *MCP* or *ACP* equilibria.

#### I. The Model

Our model will be the neoclassical two-input, two-output, two-household, two-firm

\*Cowles Foundation, Yale University, and University of Essex, Colchester, England (visiting professor at Cowles Foundation, Yale University), respectively. Research was supported in part by grants from the National Science Foundation to Yale University. We thank the participants of the Cowles Seminar and the seminar in Economic Theory at Columbia University and John Riley for their useful comments.

economy where inputs are inelastically supplied. The inputs are capital ( $K$ ) and labor ( $L$ ). The outputs are grain ( $G$ ) and electricity ( $E$ ). Each household has a utility function denoted  $U_x$  and  $U_y$ . Endowments and shareholdings in firms are given by  $(K_x, L_x)$ ,  $(K_y, L_y)$ ;  $(\theta_{xG}, \theta_{xE})$ ,  $(\theta_{yG}, \theta_{yE})$ . Each firm has a production function,  $F_G$  and  $F_E$ , or equivalently, cost functions  $c_G$  and  $c_E$ . Let  $K = K_x + K_y$  and  $L = L_x + L_y$ .

We make the same assumptions regarding firms and households as does Francis Bator in his classic 1957 expository piece on welfare economics, with one exception: we do not assume constant returns to scale, although firms are assumed to exhibit diminishing marginal rate of substitution along any isoquant, that is, the markets for inputs are competitive.

Under these assumptions, we construct the Edgeworth-Bowley box for production and the social production possibility frontier,  $PPF$ . In general, the social production possibility set is nonconvex.

Let  $P_G$  and  $P_E$  denote the prices of grain and electricity, and  $w$  and  $r$  denote the prices of labor and capital. The marginal rate of transformation ( $MRT$ ) at a point  $(\tilde{G}, \tilde{E})$  on the social production possibility frontier is simply the absolute value of the slope of the frontier at that point and will be denoted  $P_E/P_G$ . A point  $(\tilde{G}, \tilde{E})$  is said to be production efficient if it lies on this frontier.

Each point  $(\tilde{G}, \tilde{E})$  on the frontier also determines a unique point in the Edgeworth-Bowley box for production, that is, the point on the efficiency locus corresponding to the tangency of the isoquants defined by  $F_E(L_E, K_E) = \tilde{E}$  and  $F_G(L_G, K_G) = \tilde{G}$ . The slope of their common tangent line at this point will be denoted as  $w/r$  and is the marginal rate of technical substitution ( $MRTS$ ) at this point.

We shall use repeatedly that the  $MRT$  at a point  $(\tilde{G}, \tilde{E})$  is the ratio of the marginal costs; that is,  $P_E/P_G = \partial c_E(w/r, \tilde{E})/\partial E / \partial c_G(w/r, \tilde{G})/\partial G$ .

## II. Marginal Cost Pricing

As is common in this literature, we assume that each household's income at the prevailing prices is a fixed proportion of  $GNP$ . This

assumption, which is called a fixed structure of revenues, guarantees that households have positive after-tax income.

Formally, a fixed structure of revenues is defined as follows: there are fixed  $\alpha_x$  and  $\alpha_y$ , where  $\alpha_x, \alpha_y > 0$  and  $\alpha_x + \alpha_y = 1$  such that  $(K_x, L_x) = \alpha_x(K, L)$ ,  $\alpha_x = \theta_{xG} = \theta_{xE}$ ;  $(K_y, L_y) = \alpha_y(K, L)$ ,  $\alpha_y = \theta_{yG} = \theta_{yE}$ . Hence, the income of household  $x$ ,

$$\begin{aligned} I_x &= wL_x + rK_x + \theta_{xG}(p_G G - wL_G - rK_G) \\ &\quad + \theta_{xE}(p_E E - wL_E - rK_E) \\ &= \alpha_x(wL + rK) \\ &\quad + \alpha_x(p_G G + p_E E - w(L_G + L_E) \\ &\quad - r(K_G + K_E)) \\ &= \alpha_x(wL + rK) \\ &\quad + \alpha_x(p_G G + p_E E - wL - rK) \\ &= \alpha_x(p_G G + p_E E), \end{aligned}$$

if production is efficient. Similarly, the income of household  $y$ ,  $I_y = \alpha_y(p_G G + p_E E)$ . Consequently, a household's income depends only on the relative product prices and outputs of firms.

Note that each agent's income is positive and is net of the lump sum taxes necessary to cover the losses of firms producing with increasing returns. Hence, if electricity is produced with increasing returns, then  $\theta_{xE}(p_E E - wL_E - rK_E)$  is the lump sum tax imposed on household  $x$ . Another interpretation of the lump sum taxes is that the shareholdings carry unlimited liability.

A household's demand for products derive from utility maximization subject to its budget constraint:<sup>1</sup>

$$(1) \quad \partial U_i / \partial E_i / \partial U_i / \partial G_i = P_E / P_G,$$

$$(2) \quad p_E E_i + p_G G_i = \alpha_i(p_E E + p_G G) \quad i = x, y.$$

<sup>1</sup> Throughout this discussion we ignore the possibility of corner solutions. These will never occur if household's indifference curves do not cut the coordinate axes and both marginal and average costs are well behaved at zero output.

A firm's demand for factors derive from cost minimization subject to its output constraint:

$$(3) \quad \partial F_j / \partial L_j / \partial F_j / \partial K_j = w/r,$$

$$(4) \quad F_j(L_j, K_j) = j \quad j = E, G.$$

Equilibrium is defined as a set of relative prices  $P_E/P_G$  and  $w/r$ ; product demands  $E_x$ ,  $G_x$  and  $E_y$ ,  $G_y$ ; factor demands  $L_E$ ,  $K_E$  and  $L_G$ ,  $K_G$ ; and output levels  $E$  and  $G$ , such that all markets clear. That is,

Product Markets:

$$(5) \quad E_x + E_y = E, \quad (6) \quad G_x + G_y = G,$$

Factor Markets:

$$(7) \quad L_E + L_G = L, \quad (8) \quad K_E + K_G = K.$$

Because of Walras' law, equation (6) is redundant. Our final equation is the market pricing equation which gives the relationship between relative prices in the factor and product markets:

$$(9) \quad \frac{\partial c_E(w/r, E)}{\partial E} / \frac{\partial c_G(w/r, G)}{\partial G} = P_E/P_G$$

We only need two rather than three relative prices because of the fixed income distribution assumption.

Hence, a MCP equilibrium in this economy is characterized by a system of twelve independent equations in twelve unknowns. We now use a fixed-point argument to demonstrate the existence of a solution to this system of equations.

LEMMA: A continuous map,  $f(x)$ , of a compact interval of the real line into itself has a fixed point.

PROOF:

Let the interval be  $[-1, 1]$  and  $f: [-1, 1] \rightarrow [-1, 1]$ . Let  $g(x) = f(x) - x$ , then  $g(-1) \geq 0$  and  $g(1) \leq 0$ . Hence, there is some  $x \in [-1, 1]$  such that  $g(\bar{x}) = 0$ , i.e.,  $f(\bar{x}) = \bar{x}$ .

The social production possibility frontier,  $PPF$ , can be "stretched" onto a compact

interval of the real line without tearing it. More formally,  $PPF$  is homeomorphic to a compact interval of the real line. Therefore, by the lemma, any continuous map  $\Phi$  of  $PPF$  into itself will have a fixed point.

Consider the following continuous map,  $\Phi$ , of  $PPF$  into  $PPF$   $(E_1, G_1) \rightarrow P_{E_1}/P_{G_1} \rightarrow (E_2, G_2) \rightarrow (E_3, G_3)$ , where

(i)  $(E_1, G_1)$  is an arbitrary point on  $PPF$

(ii)  $P_{E_1}/P_{G_1}$  is the MRT at  $(E_1, G_1)$

(iii)  $(E_2, G_2)$  is the aggregate demand at relative product prices  $P_{E_1}/P_{G_1}$ , given production outputs  $E_1$  and  $G_1$ .

(iv)  $(E_3, G_3)$  is the intersection of the ray from the origin through  $(E_2, G_2)$  and  $PPF$ —under our assumptions on the technology, this intersection is unique.

Note that  $(E_2, G_2)$  lies on the line through the point  $(E_1, G_1)$  with slope  $-P_{E_1}/P_{G_1}$  by Walras' law. We can now prove the following theorem.

THEOREM 1: A production efficient MCP equilibrium exists, i.e., equations (1) through (9) have a solution, which is production efficient.<sup>2</sup>

PROOF:

Let  $(E^*, G^*)$  be a fixed-point of the map  $\Phi$ . At such a point  $(E_1, G_1) = (E_2, G_2) = (E_3, G_3) = (E^*, G^*)$ . Hence, demand,  $(E_2, G_2) = (E_1, G_1)$ , supply, at the relative product prices  $P_E^*/P_G^*$ , the MRT at  $(E^*, G^*)$ . The equilibrium relative prices in the factor markets,  $w^*/r^*$ , is the MRTS at the tangency of  $F_E(L_E^*, K_E^*) = E^*$  and  $F_G(L_G^*, K_G^*) = G^*$  in the Edgeworth Bowley box for production. This point of tangency gives us the equilibrium values of  $L_E^*$ ,  $K_E^*$  and  $L_G^*$ ,  $K_G^*$ ; where  $L_E^* + L_G^* = L$  and  $K_E^* + K_G^* = K$ . Finally, the household demands  $G_x^*$ ,  $E_x^*$  and  $G_y^*$ ,  $E_y^*$  total to the aggregate demand  $G^*$  and  $E^*$ . This completes the proof.

<sup>2</sup>As described, the equilibrium involves a decision by the planner, not only about how much electricity should be produced, but also about the production of grain. Each output price is then set equal to marginal cost. (Actually, all we require is that the price ratio be equal to the ratio of marginal costs.) However, profit maximization by a price-taking grain producer would also result in an output choice equating price and marginal cost so the equilibrium of this theorem remains viable with intervention only in the increasing returns to scale sector.

Note that in this model the existence of a socially inefficient *MCP* equilibrium is precluded by the assumption that inputs are fully employed. Note, also, that a *MCP* equilibrium need not be Pareto efficient despite the fact that the first-order conditions for Pareto efficiency hold at a *MCP* equilibrium. The reason is that for nonconvex production possibility sets, the first-order conditions are not sufficient to insure optimality. In a recent paper, we constructed an example of an economy, with increasing returns to scale, where all of the *MCP* equilibria are Pareto inefficient.

### III. Average Cost Pricing

Initially, we assume that only electricity is produced with increasing returns and is priced at average costs; while grain is produced with constant or decreasing returns and is priced at marginal cost. Consequently, the income of household  $x$ ,  $I_x = wL_x + rK_x + \theta_{xG}(p_G G - wL_G - rK_G)$ . Similarly, the income of household  $y$ ,  $I_y = wL_y + rK_y + \theta_{yG}(p_G G - wL_G - rK_G)$ .

Note that we do not assume a fixed schedule of revenues in this section of the paper.

The system of equations describing an *ACP* equilibrium differ from the *MCP* equilibrium equations in the following manner: Equations (1), (3), (4), (5), (6), (7) and (8) remain the same. We shall denote the new equations with primes.

$$(2') \quad P_E E_i + P_G G_i = wL_i + rK_i \\ + \theta_{iG}(P_G G - wL_G - rK_G) \quad i = x, y$$

$$(9') \quad P_G/r = \frac{\partial c_G(w/r, G)}{\partial G}$$

$$\text{and} \quad P_E/r = c_E(w/r, E)/E.$$

Equilibrium is defined as a set of relative prices  $P_E/r$ ,  $P_G/r$ , and  $w/r$ ; product demands  $E_x$ ,  $G_x$  and  $E_y$ ,  $G_y$ ; factor demands  $L_E$ ,  $K_E$  and  $L_G$ ,  $K_G$ ; and output levels  $E$  and  $G$ , such that all markets clear. Hence, an *ACP* equilibrium in this model, is characterized by a system of thirteen independent equations in thirteen unknowns.

Consider the following continuous map,  $\Psi$ , of *PPF* into *PPF*:

$$(E_1, G_1) \rightarrow (w/r) \rightarrow (P_{E_1}/r, P_{G_1}/r) \\ \rightarrow (E_2, G_2) \rightarrow (E_3, G_3),$$

where

(i)  $(E_1, G_1)$  is an arbitrary point on *PPF*,

(ii)  $w/r$  is the *MRTS* at the tangency of  $F_E(L_{E_1}, K_{E_1}) = E_1$  and  $F_G(L_{G_1}, K_{G_1}) = G_1$  in the Edgeworth-Bowley box for production,

(iii)  $P_E/r = c_E(w/r, E_1)/E_1$

(iv)  $P_G/r = \partial c_G(w/r, G_1)/\partial G$

(v)  $(E_2, G_2)$  is the aggregate demand at the relative prices  $w/r$ ,  $P_{E_1}/r$ ,  $P_{G_1}/r$  given production outputs  $E_1$  and  $G_1$ ,

(vi)  $(E_3, G_3)$  is the intersection of the ray from the origin through  $(E_2, G_2)$  and *PPF*.

**THEOREM 2:** *A production efficient ACP equilibrium exists.*

**PROOF:**

Let  $(E^*, G^*)$  be a fixed point of the map  $\Psi$ , then use the argument in the proof of Theorem (1) to complete the proof.

Finally, if both firms produce with increasing returns to scale, we require that grain is also sold at average cost:  $P_G/r = c_G(w/r, G)/G$ ; and making an obvious change in the definition of  $\Psi$ , that is, (iv) is now  $P_{G_1}/r = c_G(w/r, G_1)/G$ , we can prove the following theorem.

**THEOREM 3:** *A production efficient ACP equilibrium exists, where each firm breaks even.*

### IV. Regulation

We now review several of the standard partial equilibrium policy prescriptions for regulating a public monopoly. In the present general equilibrium model, they are simply decentralized interpretations of a *MCP* or an

ACP equilibrium, and their consistency and production efficiency are therefore assured by Theorems 1 and 2.

We consider first the policies of Oskar Lange and Abba Lerner. Lange proposed that the public monopoly, electricity ( $E$ ), be given the desired output  $E^*$  which it should produce at minimum cost and sell at average cost, subject to the prevailing relative factor prices  $w^*/r^*$ . Although this policy was put forward by Lange as an application of the marginal cost pricing principle, it is clearly only consistent with the notion of an ACP equilibrium, if electricity is produced with increasing returns to scale.

Later, Lerner modified Lange's proposal by suggesting that the desired output  $E^*$  must be sold at marginal cost if the pricing rule is to satisfy the necessary conditions for Pareto optimality. The appropriate notion of equilibrium in this case is that of a MCP equilibrium.

Turning to more current policy prescriptions, we note that the primary activity of a public utility's regulatory commission is the setting of rates, that is, prices. A common prescription is to fix the rate of return for the public monopoly; require it to meet all demand; and have it produce efficiently.

In our model, this corresponds to an ACP equilibrium where the regulator sets the price  $P_{E^*}/r$ ; requires the public monopoly to meet all demand; and to produce the demand  $E^*$  at minimum cost. In this case, the public monopoly makes normal economic profits, that is, breaks even.

Another regulatory policy, which one sees in real world markets, for example, British Rail, is that the public monopoly, electricity ( $E$ ), is first given a subsidy  $S$ , and then required to price at marginal cost subject to a break-even constraint. That is, given the prevailing factor price ratio  $w/r$  find an output  $E$  which satisfies the following equation:

$$(10) \quad c_E(w/r, E) - E \frac{\partial c_E(w/r, E)}{\partial E} = S;$$

and sell  $E$  at marginal cost.

The efficacy of this regulatory policy in a general equilibrium model reduces to a question of the existence of a subsidy  $S$ , such that the output and price of the public monopoly are consistent with the utility-maximizing behavior of households; the profit-maximizing behavior of competitive firms; and the clearing of product and factor markets.

If  $E$  is produced with decreasing marginal costs, then (10) has at most one solution. Note that if  $F_E(L_E, K_E)$ , the production function for electricity, is homogeneous of degree  $r$ , where  $r > 1$ , then  $c_E(w/r, E)$  is concave, for example,  $F_E$  is Cobb-Douglas with increasing returns to scale.

Clearly, any MCP equilibrium implicitly defines a subsidy  $S^*$  with the desired properties. Simply let  $S^* = c_E(w^*/r^*, E^*) - E^* \partial c_E(w^*/r^*, E^*) / \partial E$ , where  $w^*/r^*$  and  $E^*$  are the MCP equilibrium values of  $w/r$  and  $E$ . Note that  $S^*$  is raised through lump sum taxation.

It should be noted that all these regulatory policies lack behavioral incentives, both in the MCP or ACP interpretations, and that the development of incentive compatible variants is an important and difficult task.

## REFERENCES

- Bator, Francis M., "The Simple Analytics of Welfare Maximization," *American Economic Review*, March 1957, 47, 22-59.
- Boiteux, M., "On the Management of Public Monopolies Subject to Budget Constraints," *Journal of Economic Theory*, September 1971, 3, 219-40.
- Brown, Donald and Heal, Geoffrey, "Equity, Efficiency and Increasing Returns," *Review of Economic Studies*, 1979, 46, 571-85.
- Hotelling, Harold, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, July 1938, 6, 242-69.
- Lange, Oskar and Taylor, Fred M., *On the Economic Theory of Socialism*, New York, McGraw-Hill, 1964.

# The Present Direction of the FCC: An Appraisal

By NINA W. CORNELL AND DOUGLAS W. WEBBINK\*

Since at least the Ford Administration, deregulation has been a politically popular slogan. Over the past ten years there has been a growing recognition that dismantling the economic regulation of various sectors—airlines, trucking, communications, and the like—requires simultaneous or prior dismantling of the barriers to entry and competition. This recognition was translated into action during the Ford and Carter Administrations when the chairmen of the CAB, the ICC, and the FCC were trying to convince a majority of their fellow commissioners or board members (and the relevant members of Congress) to adopt promarket, deregulatory positions. During those periods, each agency made significant progress in opening up entry to new firms and new service offerings, and in reducing the degree of detailed intervention in the daily affairs of the regulated companies.

Since the advent of the Reagan Administration, however, the momentum pushing open entry and promoting competition in communications has definitely been blunted, if not entirely broken. As a result, it may not be politically feasible to let the market, rather than the government, serve as the arbiter of the public interest in communications.

Because of some basic differences between the industries, opening entry to communications markets requires different policy actions from those required to allow entry into most transportation sectors. Communications is characterized by two features not present in airlines and trucking: the use of the frequency spectrum and the market power of certain firms caused by the presence of a very significant proportion of sunk vs. variable costs. The use of the spectrum realistically means that government will determine basic entry possibilities into the foreseeable future by the way that it allocates spectrum, a process similar to making land zoning deci-

sions. Having a high proportion of sunk vs. variable costs means that antitrust-type issues will remain important. While this latter condition is seen primarily as applying to common carrier communications at this time, the rapid growth and spread of cable television systems will make this condition applicable to broadcasting markets in the future.

Given these features, further movement towards a truly promarket, deregulatory policy in communications requires that spectrum allocation decisions be made with the explicit goals of encouraging entry, increasing competition, and decreasing the market power of currently dominant firms. In particular, spectrum needs to be allocated in a way that assures that spectrum is always available to permit new service offerings in competition with those being offered over wire systems.

In this respect, the Fowler FCC has deviated from its most recent predecessors. The Fowler Commission seems to be concerned more with removing some, but by no means all, restrictions on existing firms than with encouraging competition and new entry into the industry. It has not been willing to expand the ability of new firms to offer existing services, where competition could come about most rapidly. Rather, to the extent that it has considered allowing entry at all, it has focused its attention on creating new services that are either secondary to existing services, are structured so they can only supplement and not compete with existing services, or are able to compete only far in the future. Recent FCC activities illustrate these deficiencies.

## I. Examples of FCC Failures to Open Entry

When President Reagan took office, the FCC had a number of proceedings underway that would have permitted use of its frequency allocation (or allotment) powers to expand entry into broadcasting. The Fowler FCC

\*Cornell, Pelcovits & Brenner Economists Inc.

has acted on only some of these. In 1982, the FCC finally approved a new low-power television service. The FCC has also approved interim rules for direct broadcast satellites, and has taken the first steps towards expanding the AM radio band.

These actions, while they may increase competition sometime in the future, are much more limited in their likely ultimate impact than other steps the Commission could have taken. Many other proposed actions that would have expanded entry more effectively have been slowed or stopped completely. The Commission activities affecting both radio and video services offer many examples.

The FCC has had a number of options for expanding the number of AM radio stations, including reducing the frequency per station within the band, and expanding the AM band. The first option could have led to a more rapid increase in competition, as well as to put the new stations on a more equal footing with the existing stations. In the summer of 1981, the Commission voted not to reduce the AM radio station channel spacing from 10 to 9 kilohertz, even though the reduced spacing could have allowed a significant number of new stations to go on the air. (All the rest of the world except the Western Hemisphere uses the reduced frequency spacing.) Although its major reason was a claim of technical difficulties if the shift were made, the Commission also pointed to a forthcoming expansion of the AM band as sufficient to provide for increased entry into AM broadcasting.

That expansion, however, lies well in the future. Although the expansion of the AM band was decided in principle at the World Administrative Radio Conference in 1979, the steps necessary to begin accepting license applications will not be complete until the middle of the decade, after another international conference. Even when licenses are finally made available, in order to have any listeners, new station owners will have to convince the public to buy new radios, because existing radios do not receive the frequencies involved. The Commission could have made life far easier for new entrants on the AM band if it had chosen to reduce the channel spacing.

Perhaps more serious, given the emergence of FM as the dominant medium in the radio market, is the failure of the FCC to move on *any* entry expansion in the FM band. In 1980, the Commission proposed establishing two new classes of FM stations, which would allow many additional FM stations to go on the air, but to date it has not yet finalized those rules, much less granted any licenses for those new stations.

In summary, looking at the actions on radio in the last two years, the FCC has backed away from making any changes that would bring in new competitors any time soon. Moreover, the potential new competitors being permitted some day in the distant future will face two big hurdles to becoming actual competitors: winning a comparative hearing if any other firm wants the license; and then persuading the public to invest in new AM radios. Existing licensees have little to fear from such a competitive threat.

Similarly, the current FCC has chosen extremely slow ways to expand video services. As was true with radio, the FCC in 1981 had a number of ongoing proceedings that could have brought in a number of new video services that would offer real competition to the existing firms. New video distribution possibilities included allowing new VHF stations that could be engineered in between existing ones (VHF drop-ins), low-power television stations in both the VHF and UHF bands, expansion of the multipoint distribution service (*MDS*)—another method of distributing pay television programming besides subscription television in the UHF and VHF bands and pay cable—and authorization of a direct broadcast satellite service (*DBS*).

The Commission to date has authorized only a low-power television service and *DBS*. While both offer promise of future competition to the existing television and cable television services, the promise will not be redeemed any time soon. Indeed, in the case of the low-power television service, the FCC set up its license processing procedures to delay, if not prevent, the rapid growth of new stations as a competitive alternative to existing stations.

Those who first proposed low-power TV service saw it as an exciting, new, low-cost

service that would allow many new and diverse entrants into the industry. Although the FCC has authorized the service, only a small number of applications (a little more than 100) have been granted despite a huge backlog of applications (over 7,000), many of which appear to be mutually exclusive. The new law authorizing the Commission to use lotteries instead of slow and costly comparative hearings to choose among competing applicants may speed up the licensing process, but even so the process will take years.

The delay in offering effective competition due to time-consuming licensing practices has been heightened by a policy decision on licensing priorities. The FCC has announced that it intends to process applications for stations in the most rural parts of the country first, while saving for last applications to serve the urban areas where most of the potential audience lives. As low-power stations are required to give way to any existing or new full-powered stations in their area, there may never be low-power service in some cities if would-be low-power applicants fear full-power applications may be coming soon.

The *DBS* likewise offers only a promise of possible future competition. Although interim rules for *DBS* systems have been approved, no applicant will be assigned specific frequencies and orbital slots until after the Regional Administrative Radio Conference (RARC) in 1983, and it will be at least several years after 1983 before any direct broadcasting satellites can operate. Indeed, there is currently no way to know whether mutually exclusive applications exist in *DBS*. That will depend on the outcome of the 1983 RARC.

Again, the Commission could have chosen other, more rapid and effective means of expanding competition in the video market. In September 1980, the Commission approved the idea of allowing reduced-power VHF drop-in stations to be located between full-power VHF stations in four specific markets, but the Commission has yet to grant a single license for a drop-in. Nor has it moved to allow, even in principle, such drop-ins in many other places in the country where they could be built without causing unaccep-

table interference to existing stations. Because VHF drop-in stations would operate at much higher power levels than low-power TV stations and, thus, could cover a much higher potential audience, they pose a much larger competitive threat to existing stations.

Similarly, if the Commission wanted to encourage entry and increased competition to cable television with both its multichannel capability and its pay television services, it could adopt the proposal to increase the number of *MDS* channels available in each location. Although the Commission issued a number of proposals in 1980 to make additional channels available to *MDS*, the Commission has taken no further action other than accepting additional comments.

The examples above all relate to broadcast services, but the same story emerges from an examination of FCC actions affecting common carriers. As was the case with broadcast services, at the start of the Reagan Administration the FCC had a number of ongoing proceedings to allocate spectrum to several new services that could compete with some of the common carrier services offered by AT&T (the proceedings involving the new Digital Electronic Message Service, or *DEMS*, and the new cellular mobile radio service).

To date, the Fowler FCC has not used those proceedings to increase the effective competition to AT&T. While both of the new services have been created, *DEMS* is limited to certified common carriers, meaning in practice that many fewer firms will be interested. Cellular mobile radio service has been set up with a guarantee that half of the available spectrum will be granted to the existing wireline common carriers, thus making it almost impossible for cellular service ever to compete directly with wireline local distribution.

## II. New Communications Legislation is Needed

The Fowler FCC's record has occurred despite the administration's rhetoric in favor of deregulation and letting markets work. It also flies in the face of continued congres-

sional interest in seeing new services made available to the public. It is a sobering reminder that the mandates facing regulatory agencies are very flexible indeed. Given the track record of the present FCC, if pro-market policies are desired, new legislation will be required, although legislation cannot guarantee that this or any other FCC will follow appropriate policies.

Among other needed changes to improve the competitive record of communications markets, new, pro-market communications legislation would make two changes in the current law. First, and most important, it would require the FCC to make spectrum allocation decisions based on antitrust-type concerns. In other words, the Commission should be required to favor allocating spectrum to new services that would compete with present services. In particular, the Commission should be required to allocate spectrum in such a way that spectrum-based services could always be offered in competition with wire-based services. This means ensuring that over-the-air television spectrum be kept available as a competitive alternative to cable television. It also means changing the philosophy that led to guaranteeing half of the cellular radio frequencies to the existing wireline carriers. Given present costs, cellular mobile radio is unlikely to be competitive with wireline distribution, at least for most users, but the future possibility of such competition should not have been foreclosed.

Second, as part of the mandate for increased competition, the Commission also should be forbidden from using the competitive impact on existing firms as a reason for not allowing new entry. Today new applicants and new services bear the burden of

showing that their proposed services are wanted by the public and would not hurt existing firms. Congress should instruct the Commission to ignore the impact on existing firms and to require that those who would prevent entry disprove benefits to the public.

Two other changes in the Communications Act would reinforce the procompetitive nature of the changes advocated above. First, comparative hearings as a means of assigning licenses to specific firms should be abolished and replaced by lotteries or auctions. Further, the remaining restrictions on trading or selling licenses once granted should be repealed.

Finally, the Commission must be granted the power to exempt any firm or set of firms from regulation wherever the basic goals of the Act would be better served by such exemption. In the common carrier area that would mean that firms who do not have substantial market power or who are rapidly losing market power should not be subject to any part of Title II. In other words, "competitive common carriers" should not be subject to *any* of the traditional common carrier requirements: duty to serve, entry and exit control, tariff filings, nondiscrimination, allowed rate of return, and the like. Broadcasting firms that face significant competition should answer to their listeners or viewers, not to the FCC on what programming is "in the public interest."

Without these changes, the Fowler Commission seems likely to continue as it has begun: engaging in only minimal deregulation and permitting only entry that is relatively painless to those who are already in the market. Neither serves well either taxpayers or the consuming public.

# The First Step in Bank Deregulation: What About the FDIC?

By JOHN H. KAREKEN\*

If insuring creditors of commercial banks in the way they have been insured since mid-1933 is justified, then so is regulation of so-called insured banks, those with creditors insured by the Federal Deposit Insurance Corporation (FDIC). In creating the FDIC, the Congress mandated a pricing policy: all banks with FDIC-insured creditors were to be charged alike; more particularly, the FDIC was not to charge insured banks according to the riskiness of their respective balance sheets. Nor has it ever. Yet, with an insurance premium that is constant across balance sheets, there is an incentive for risk taking. And thus, unless insured banks are to be as risky as profit maximization dictates, they must, one way or another, be effectively regulated; they must, that is, be limited by regulation to appropriately risky balance sheets.

It does not follow that U.S. bank regulatory policy of the years since 1933 is beyond criticism. For example, we must wonder about the geographical restrictions imposed under the McFadden Act and the Douglas Amendment to the Bank Holding Company Act. But it does follow that if banks with FDIC-insured creditors are to be made entirely free, except perhaps of reserve requirements, or even largely free, then it is necessary to either close up the FDIC or, if doing that seems unwise, change FDIC policy.

There has already been some deregulation. Most importantly, Regulation Q has been made much less effective than it was; and evidently it has been marked to become, very soon, a thing of the past. So far, however, beyond deregulating, the Congress has not bestirred itself. The FDIC is still occupying its Washington corner. Although FDIC officials have hinted at change, its policy is still by and large what it was. And the Congress,

being ever so respectful of the consumer lobby, may never want to do anything. Nevertheless, in this paper I consider various things it might do: namely, close down the FDIC, but at the same time impose a new valuation rule for bank portfolios; change FDIC pricing policy; and, lastly, without doing anything else, simply close down the FDIC.

I do not end up by saying what, as I believe, the Congress ought to do. My purpose is only to determine, as best I am able, which of those several apparent congressional options of mine are in reality feasible. I would add, however, that the Congress, if bent on deregulating banks or obliging the regulatory agencies, has more to do than decide what to do about the FDIC. It seems also to be bent on deregulating or allowing the deregulation of savings and loan associations. There has already been more deregulation of savings and loan associations than of banks. So the Congress has also to decide what to do about the Federal Savings and Loan Insurance Corporation (FSLIC). Fortunately, to explore what it might do about the FDIC is perforce to explore what it might do about the FSLIC.

## I. Why Insure Bank Creditors?

The desirability of protecting bank creditors against default has been argued in different ways. Thus, there are those who believe that a bank account, quite free of default risk (up to some real dollar limit), ought to be a kind of birthright of every U.S. citizen. And those who believe that argue for government-insured bank accounts. There is an alternative: change the denomination of, say, the three-month Treasury bill from \$10,000 to \$100 or even \$5. The greater borrowing cost must appear insignificant when matched against the cost of regulating banks. There are, however (or so the argument goes), many individuals who are hope-

\*Professor of finance and economics, School of Management, University of Minnesota. I have benefited substantially from conversations with Neil Wallace and Carter Golembe.

less as risk appraisers; were Treasury bills available in small denomination, most would likely not buy them. But nearly all of us have bank (checking) accounts. So insuring those accounts does the trick.

The birthright argument, shaped in the heart, is difficult to deal with, and I content myself with noting the obvious: it does not, for example, justify a \$100,000 insurance limit. No one thinking only of "the farmer, the mechanic and the laborer" would insist on so generous a birthright. Possibly we should all have a portfolio option (other than holding currency, which is inconvenient) that is free of default risk. But if providing a birthright, appropriately modest, for every citizen is the justification for insuring bank accounts or creditors, then FDIC policy should be drastically revised.

If, alternatively, the justification is that in the United States we have an "inherently unstable" banking industry, one prone to bank runs which are frightfully costly, then the policy of the FDIC appears much more consistent with its purpose. And that justification, a banking industry prone to costly bank runs, is a part of the conventional wisdom of university economists. According to that conventional wisdom, government insurance is not necessary for preventing bank runs. A central bank, acting as the lender of last resort, can prevent them. But our central bank, the Federal Reserve, is not to be trusted. In the early 1930's it let us down badly; lacking either in will or in understanding, it simply watched as thousands of banks failed and, in so doing, allowed an ordinary recession to grow into an extremely severe and traumatic depression. So, whatever theory may say, there is then, as a practical matter, a need for an insuring agency, an FDIC.

Among the benighted, the bank run is invariably portrayed as the doing of a frenzied mob. It may, however, be the doing of (in the economist's sense) rational individuals. Think of someone who has just gotten word of the failure of a bank, not his own. If not perfectly informed about the balance sheet of his bank, he might reasonably revise his estimate of his default risk and, depending on the cost, turn into currency whatever

is due him on demand. That could easily be the optimal strategy.

We must be skeptical of any happening alleged to result from a large number of individuals behaving in a silly way. It is thus not unimportant that a bank run can occur in an economy populated exclusively with rational individuals. If that were not so, the university economists' conventional justification for insuring bank creditors would be less appealing than it is.

And were it not for very recent a contribution by Douglas Diamond and Philip Dybvig, the conventional justification would perhaps be less appealing than it is. For most university economists, there has long been a disposition to just accept that the failure of many banks has to be costly. Is not chaos always costly? To Milton Friedman's credit, he recognized that a happening being costly does not suffice. A strike may be in some sense costly. So should the federal government, with soldiers at its disposal, ban all strikes? But in *A Program for Monetary Stability*, which is where Friedman takes up the question of whether government should be concerned about how risky banks are, one finds only brief unargued statements (see pp. 4-8): keeping banks free of default risk amounts only to assuring contract compliance; and, what is supposedly the clincher, a wave of bank failures has third-party effects. In any default, though, there is renigging. And if a bank run has third-party effects, what about the failure of several or more large money market funds?

In contrast with Friedman and others, Diamond and Dybvig are, mercifully, entirely explicit. In their world, elegant in its simplicity, a bank run lessens economic well-being. And that is an implication, one with definite meaning. Yet, the Diamond-Dybvig world is one without "money."

A critical characteristic of that world is an "illiquid" production technology: one unit of investment at time  $t$  yields one unit of consumption at time  $t+1$  or, with no consumption at time  $t+1$ ,  $R > 1$  units of consumption at time  $t+2$ . Another critical characteristic is the existence of private information. So in the Diamond-Dybvig world, there can be no ordinary risk-sharing con-

tracts arranged in a private insurance market. And if a "bank" exists, then the competitive equilibrium consumption allocation is not always Pareto optimal. (Diamond and Dybvig do not refer to the literature on why financial intermediaries exist, but have, I believe, made an important contribution to it.) With no bank run, the equilibrium allocation is Pareto superior to the competitive equilibrium allocation that obtains when there is no bank; and if there is a run, then it is not. Thus do Diamond and Dybvig find advantage in tax-financed government insurance for bank creditors: or, to be more precise, if there is no uncertainty, in a prohibition against withdrawals; and if there is uncertainty, in tax-financed government insurance, which then yields a Pareto optimal allocation.

That Diamond and Dybvig have said all that wants saying is doubtful. Indeed, I wonder if it is so that in their world ordinary risk-sharing contracts cannot in effect be arranged. The Diamond-Dybvig bank may be nothing more than a substitute for a private insurance market. I could, however, be dead wrong, and if so, then thanks to Diamond and Dybvig, the conventional justification for insuring bank creditors is more appealing than it was. It is easier than it was to accept, if still only provisionally, that government insurance for bank creditors serves the worthy purpose of preventing welfare-reducing bank runs.

## II. An Alternative to Government Insurance?

Evidently, providing birthrights and preventing bank runs are very different purposes. Nor is it possible, if preventing bank runs is the purpose, to leave some bank creditors uninsured, even in the hope that the uninsured creditors, having to be ever watchful, will keep banks from becoming too risky. Banks do have to be kept from becoming too risky. But coinsuring—that is a word used in official circles to describe the policy of not insuring all creditors in full—is not a feasible way. To prevent bank runs, it is required that all creditors, the owners of large-denomination CDs included, be entirely insured. (It is also required that, in the

instance of failure, there be no netting; an owner of a large-denomination CD might flee if there were some risk of it being used, in whole or in part, to pay off a loan.) Owners of large-denomination CDs might decide, more or less together, not to renew their loans, and their doing that would approximate a bank run.

The FDIC has often been criticized for having paid out (or off) only rarely. Actually, from 1945 through 1981, there were 67 payouts. It is true, though, that in almost every instance of the failure of a bank with deposit liabilities of \$25 million or more, the FDIC managed to arrange a purchase and assumption. And it has carried on in 1982. That is what many critics, mostly university economists, do not like, no doubt because there is more of an incentive for banks to be risky when all creditors are insured in full than when some are virtually uninsured. But if the purpose of insuring bank creditors is to prevent bank runs, then the FDIC has behaved well. It is to be criticized not for having paid out only when "small" banks have failed, but for having done so at all. With the way the Federal Deposit Insurance Act reads, it takes purchase and assumption deals to insure all bank creditors in full.

It also appears that the regulatory authorities (among them, the Federal Reserve) did well in their handling of the failure of Franklin National Bank. As will be recalled, the Comptroller of the Currency—in a manner of speaking, financed by the Federal Reserve—put off declaring Franklin National insolvent until nearly all owners of large-denomination CDs had been paid off by the bank. In contrast, the FDIC may have made rather a serious mistake in its handling of the failure of Penn Square National Bank. It did a pay-out, and some owners of large-denomination CDs suffered capital losses (or will end up having done so). In effect, then, the FDIC announced that not all bank creditors are insured in full. It was common knowledge, though, that Penn Square officers had been behaving peculiarly. And had they not been, perhaps the FDIC would not have dared a pay-out. In any event, it was easy to conclude that the failure of Penn Square was special and therefore

that what the FDIC did was too. So possibly bank creditors heard only a whisper from the FDIC, not a thunderous warning. Nevertheless, there is the danger that at some time in the future, bank creditors will recall only that some Penn Square creditors lost out.

To give the FDIC's critics their due, it is right that insuring all bank creditors in full invites a misallocation of resources, more of a misallocation than when some creditors are insured only in part or not at all. But whether all or only some creditors are insured in full, there may not have to be any actual misallocation. In theory, regulation can keep banks from becoming excessively risky. Then, too, there is the intriguing possibility of pricing FDIC insurance properly. That is something to be considered, and I will, but only after having suggested an alternative to government insurance as a way of preventing bank runs.

From the beginning, it seems, U.S. banks have issued bond-like or state-independent fixed-dollar claims. The promise has always been to pay back, regardless of the state of the world, a fixed number of dollars. That has been little remarked on; and I am puzzled as to why, for that U.S. banks have issued fixed-dollar claims, and, at the same time, held risky assets, is why the U.S. banking industry has been "inherently unstable" or prone to runs.

The issuing of fixed-dollar claims is not alone a necessary condition for a banking industry being prone to runs. But, despite assertions to the contrary, neither is banks holding only fractional reserves. For a banking industry to be run-prone, it is required that banks have fixed-dollar claims and, if demand claims, assets other than currency. More generally, it is required that banks be unhedged (in the sense of having a particular mismatch of asset and liability maturities), or be subject to some interest rate risk.

It might be argued that something more is required: that banks must follow the rule of "first come-first serve" in dealing with creditors. That is right; they must. But their doing so would appear to be inevitable. Until a bank has sold off all its assets or, in anticipation of demise, suspended convertibility, it has no alternative. What would it mean for a

bank to tell a creditor with, for instance, a demand claim to wait for payment until it (the bank) has rounded up all of its other such creditors? So, in my estimation, I am entitled to regard banks having fixed-dollar claims and risky portfolios as being necessary for a run-prone banking industry.

Why fixed-dollar claims and risky portfolios are together necessary is easily explained. If banks issued equity-like claims or had riskless portfolios, no demand creditor would ever see advantage in being first in line—or see being first in line as a way of getting what he is owed, but more than his "fair share," more than each of all other creditors will be able to get.

It is thus apparent how a banking industry might be made run proof. Requiring riskless portfolios is one possibility. In practice, however, doing that could be quite a task; for what a riskless portfolio is depends on the composition of liabilities. The other possibility, the one I had in mind, is to require banks to value their assets using current market prices. That is what some mutual funds do, although they do not have commercial loans, which may be difficult to value at market prices. Nor is there any sting in the objection that, if forced to value assets at market prices, banks would be coerced. The costliness of a bank run may justify some coercion. And to require banks to value assets at market prices is no more coercive than requiring them to value assets at purchase prices. Or than requiring banks with national charters to have FDIC insurance. Or than requiring them, as Friedman at one time would, to hold portfolios of currency against certain types of liabilities.

I started out by saying that, thanks to Diamond and Dybvig, it may be easier now than it was to see a bank run as costly and therefore to take runs seriously. But in suggesting how a banking industry might be made run proof, I used an "explanation" for bank runs that does not apply in the Diamond and Dybvig world. In that world, production of consumption goods is riskless, and underlying my explanation is an assumption that such production is risky. Some bank suffers a capital loss, having made a loan to, say, a company the management of

which made an unfortunate investment decision. With that loss becoming known, creditors of other banks begin to wonder. And if in wondering they become scared enough, a run results.

If I have been inconsistent, that is because I want an explicitly worked-out cost of a bank run and also, what Diamond and Dybvig may very well not have provided, an explanation for bank runs. And so far, no one has done an analysis of bank runs on the assumption, to me more plausible than the alternative, that production of consumption goods is stochastic, nor provided a justification of government insurance for bank creditors, a justification of the sort provided by Diamond and Dybvig, but, again, on the assumption of stochastic production. Whether providing the analysis and the justification (should there be one) will be easy, I do not know.

### III. Proper Pricing of FDIC Insurance

Among interested university economists, it was once a widely shared view that banks ought simply to be deregulated, although perhaps not completely. Supposedly, even with only some deregulation (for example, doing away with Regulation Q), there would be a gain in economic efficiency and thereby in economic welfare. Understanding has increased, though, and now, I suspect, many of the formerly naive would favor more or less complete deregulation, but only after the FDIC has substituted a new pricing policy for that of the years since 1933. Of course, before there can be such a substitution, the Congress will have to act; and if there have been hints of a greater awareness, chances are still that it will be a while in allowing the FDIC to charge banks different premiums, depending on their respective balance sheets. But whether a policy prescription makes economic sense is not to be determined by how the Congress is inclined. My concern is that we, economists and officials of the regulatory agencies, may not know enough to fashion a pricing policy that can with great confidence be substituted for all present-day banking laws and regulations.

Some, I fear, believe that a very simple pricing policy will suffice. Banks might be made to pay more or less, depending on their respective capital asset ratios. Having such a policy would be particularly unfortunate, for more capital does not by itself make for a less risky balance sheet; and we do have to keep in mind that what we are considering is the substitution of a pricing policy for all or nearly all of the laws and regulations currently applying to banks. The more important point, though, is that there are various kinds of risks. So the bank premium must be made to depend on many variables: among them, the riskiness of the loan portfolio (which is not to be measured just by the proportion of problem loans); how different asset and liability maturities are; and to mention but one more, the kinds of non-banking activities being engaged in. And do we know how all the various industries of the world economy rank as credit risks? Or what the best way is of measuring how unhedged a bank is? I rather think not.

And there are other questions. Suppose bank *A* acquires a portfolio that in the judgment of the FDIC is much riskier than the portfolio of bank *B*. It may have a greater proportion of loans to risky industries. Presumably, bank *A* should pay more for its insurance than bank *B*, although how much more is something else again. But suppose that with the passage of time bank *A* comes to have a greater proportion of nonperforming loans. It is to be charged still more than bank *B* (enough more perhaps so that its future becomes uncertain)? Many automobile insurance companies increase the premiums of those who have just had accidents. They apparently feel that going by who have accidents and who do not, they can refine their risk classifications. So it may be that bank *A*, after having experienced an increase in bad loans, should see the differential between its premium and that of bank *B* increased further. The increase in bad loans may be taken as proof of poor management. What, though, if many other banks have also experienced increases in their bad loans? Then what is to be said of the management of bank *A*?

I could be quite wrong; there may be easy answers, some already known, to the question I have posed and the many others that I have not. My guess, however, is that there is much hard thinking to be done and, with that thinking, a very considerable amount of statistical analysis.

#### IV. Doing Without Insurance for Bank Creditors

There remains one option to be considered: doing without the FDIC. When Diamond and Dybvig have written a second paper, or others (borrowing from the first Diamond-Dybvig paper) have written it for them, we may have to forget that option. And were it not for one possibility, suggested by the debate of the years just before the Federal Deposit Insurance Act was signed into law, it might seem much too risky. In that debate, many of the largest of the then-existing banks were joined in opposition to government-provided insurance for bank creditors, and ranked against them were the much smaller but ever so much more numerous independent banks. They, it seems, knew what was necessary for their continued survival, and that they are still around, many of them, is likely due more to the existence of the FDIC than to antibranching laws. But then it is a reasonable conjecture that if the Congress did away with the FDIC, the U.S. banking industry would undergo a dramatic reorganization and come to approximate

much better than at present the banking industries of the other highly industrialized countries. And if I am right that the United States has experienced many more bank runs, local and countrywide, than have those other countries, then that is a decidedly interesting possibility.

#### V. Conclusion

I end with what may be the essential point: that before starting in on deregulation of banks, the Congress should have decided what to do about the FDIC; or, since it is not too late, that it should decide before proceeding further with deregulation. To deregulate further and never do anything about the FDIC would be to invite a crash. Of course, what the Congress should do about the FDIC is something else again. With some small probability, it could be best to close up the FDIC, or keep it open as a guarantor of birthrights, but require banks to value their assets differently than they do at present and to change the promises made to creditors.

#### REFERENCES

- Diamond, Douglas and Dybvig, Philip, "Bank Runs, Deposit Insurance and Liquidity," *Journal of Political Economy*, 1983 forthcoming.
- Friedman, Milton, *A Program for Monetary Stability*, New York: Fordham Press, 1980.

## *R&D AND PRODUCTIVITY INCREASES*

### Technological Change and Market Structure: An Empirical Study

*By* EDWIN MANSFIELD\*

It has long been recognized that technological change is one of the major forces influencing an industry's market structure. Karl Marx stressed this fact over a century ago. Like Marx, many economists, including Arthur Burns (1936) and John Kenneth Galbraith (1967), have been convinced that technological change tends to increase plant sizes and the level of industrial concentration. Others, like John Blair (1972), have argued that, although such a trend existed in the past, it has been reversed since World War II because of a fundamental change in the nature of technological advance whereby centralizing technologies have been displaced and superceded by decentralizing technologies.

In recent years, there has been a revival of interest in the effects of technological change on market structure. Richard Nelson and Sidney Winter (1978) have formulated a computer model and Richard Levin (1980) has estimated an econometric model, both aimed at representing these effects. In general, these models seem to suggest that a relatively rapid rate of technological change in a particular industry is likely to result in a relatively high level of concentration. However, these authors are careful to point out that their results are preliminary and tentative.

\*University of Pennsylvania. My research was supported by a grant from the National Science Foundation. I am grateful to the Foundation, as well as to the 34 firms that provided essential information used here. Preliminary versions of parts of this paper were presented at the Conference on *R and D*, Patents, and Productivity held by the National Bureau of Economic Research in 1981, and in invited lectures at the International Institute of Management in Berlin and the University of Louvain in 1982. A more complete version of this paper is available on request from the author.

Although the effects of technological change on market structure are of fundamental importance to both economic analysis and public policy, it is surprising how little systematic study has been devoted to them. We know little or nothing about the effects of the various process and product innovations that have occurred in recent years in various industries. And we have very little information concerning the relationship between the rate of technological change in a particular industry and the changes in the industry's market structure. My purpose in this paper is to try to begin filling these notable gaps.

#### **I. Effects of Major Process Innovations on Minimum Efficient Scale of Plant**

As a first step toward testing Blair's hypothesis that, since World War II, fewer innovations have tended to increase the minimum efficient scale of plant than in the past, I obtained data regarding the proportion of major process innovations in the chemical, petroleum, and steel industries that have resulted in increases in minimum efficient scale of plant. To obtain these data, a sample of innovations was drawn at random from published lists of the major new processes in each of these industries since about 1920. (See my 1968 book, my 1977 book with others, and Ralph Landau, 1980.) Nine chemical firms, 12 petroleum firms, and 4 steel firms agreed to indicate the effect of each innovation on the minimum efficient scale of plant. For 35 of the innovations, the firms (or more accurately, their highest-level engineers) were unanimous (or virtually so) in their evaluation of the direction of the innovation's effect. Also, where possible, this

TABLE 1—PERCENTAGE DISTRIBUTION OF MAJOR NEW PROCESSES BY EFFECT ON MINIMUM EFFICIENT SCALE OF PLANT, AND OF MAJOR NEW PRODUCTS BY EFFECT ON FOUR-FIRM CONCENTRATION RATIO

Effect of Process or Product	Chemicals (1929–76)	Drugs (1947–78)	Petroleum (1919–76)	Steel (1919–60)
Percentage Distribution of Major New Processes, by Effect on Minimum Efficient Scale of Plant				
Increase	92	— <sup>a</sup>	75	43
No effect	8	— <sup>a</sup>	25	43
Decrease	0	— <sup>a</sup>	0	14
Total	100	—	100	100
Percentage Distribution of Major New Products, by Effect on Four-Firm Concentration Ratio				
Increase	43	17	60	43
No effect	29	8	40	57
Decrease	29	75	0	0
Total	100 <sup>b</sup>	100	100	100

<sup>a</sup>Process innovations in the drug industry are excluded. The emphasis of pharmaceutical R & D is on new products, and lists of new processes in drugs have not been published.

<sup>b</sup>Because of rounding errors, figures do not sum to total.

evaluation was checked against published studies.

In the chemical and petroleum industries, the bulk of these process innovations resulted in increases in minimum efficient scale of plant (Table 1). In steel, only about half of these process innovations resulted in such increases, but most of the rest had little or no effect on minimum efficient scale. Thus, in all three industries, scale-increasing innovations far outnumbered scale-decreasing innovations. And if the innovations were weighted by a measure of their importance, the results would be the same. Moreover, it is very unlikely that the preponderance of scale-increasing innovations is due merely to sampling errors. Based on the usual statistical procedures, the probability is more than 0.95 that scale-increasing innovations outnumber scale-decreasing innovations by at least 3 to 1 in these industries. These results are in accord with the observed changes in minimum efficient scale of plant. There appear to have been considerable increases in minimum efficient scale of plant in all of these industries during the relevant period.

In what ways do the characteristics of the scale-increasing process innovations differ from the others? Perhaps because they may

be more likely than other process innovations to entail the construction of a new plant or the major overhaul of an old plant, scale-increasing process innovations seem to require larger investments by users than other process innovations, and (partly for this reason) they seem to be more likely than other process innovations to be introduced initially by one of the industry's four largest firms. Further, they are much more likely than other innovations to be invented by the innovator, which is explained partly by the fact that their innovators are relatively likely to be among the largest firms. (For example, in the chemical industry, relatively small firms seem to be less likely than the largest firms to invent their own innovations; they rely more heavily on engineering firms and foreign sources of technology. See my 1977 study with others.) But with regard to their profitability to users and their rates of diffusion, scale-increasing process innovations do not seem to differ significantly from other process innovations.

To test Blair's hypothesis, I compared the proportion of innovations introduced after 1950 that resulted in an increase in minimum efficient scale of plant with the proportion introduced before or during 1950 that did so.

Contrary to Blair's hypothesis, the proportion was higher, not lower, in the later period. Of course, these results pertain only to three industries, and the situation may be different elsewhere. But in these industries at least, the data do not seem to support Blair's contention.

## II. Effects of Major Product Innovations on the Four-Firm Concentration Ratio

The bulk of firms' research and development is directed at new products, not new processes. To learn about the effects of new products on concentration, I obtained data regarding the proportion of major product innovations in the chemical, drug, petroleum, and steel industries that have resulted in increases in the four-firm concentration ratio. To obtain these data, a sample of innovations was drawn at random from published lists of the major new products in each of these industries. (See my 1968 book, my studies with others, 1971; 1977, and David Schwartzman, 1976.)

When a major new product is introduced, it generally competes with existing products. Depending on how the relevant market is defined, the new product may increase or decrease concentration. The definition of the relevant market (and in some cases, the choice of which market is most important or typical) is a thorny task requiring an intimate and detailed knowledge of the new product's characteristics and its relationships to existing products. To carry out this task, I turned for help to the firms in each industry. Nine chemical firms, 9 drug firms, 12 petroleum firms, and 4 steel firms agreed to define the relevant market and to indicate the effect of each innovation on the four-firm concentration ratio in that market. For 31 of the innovations, the firms (or more accurately, their market research and economics staffs) were unanimous (or virtually so) in their evaluation of the direction of the innovation's effect. Also, where possible, this evaluation was checked against published studies.

In the petroleum and steel industries, the concentration-increasing product innovations greatly outnumbered the concentration-decreasing product innovations. But in

the chemical industry, there were almost as many concentration-decreasing innovations as concentration-increasing innovations; and in the drug industry, the concentration-decreasing innovations outnumbered the concentration-increasing innovations. Based on the available data, there is no evidence in these industries that concentration-increasing innovations were more important, on the average, than concentration-decreasing innovations. Moreover, sampling errors are very unlikely to have been responsible for concentration-decreasing innovations being a substantial percentage of the total in the drug and chemical industries combined. If concentration-decreasing innovations were much fewer than concentration-increasing innovations in these industries combined, the probability that my results would have occurred is less than 0.04.

These results are noteworthy, given the common tendency among economists to view technological change as a concentration-increasing force. In some major industries, it appears that concentration-decreasing innovations are a very substantial proportion of the total. To see how consistent my data are with observed changes in concentration in these industries, I regressed the change in the four-firm concentration ratio in each industry in 1947-58 and 1958-67 on the percentage of the industry's product innovations (during the relevant period) that increased the four-firm concentration ratio.<sup>1</sup> As might be expected, they seem to be directly related (although the correlation coefficient is significant only at the 0.10 level). The correlation is only moderate ( $r = .51$ ), but this reflects the fact that many factors other than product innovation affect concentration ratios (and that my data are crude and contain sampling errors).

To get a better idea of why the percentage of concentration-increasing product innovations in the drug industry was relatively low,

<sup>1</sup>Of course, a simple count of innovations does not give an unambiguous signal of their net effect on concentration. Even if more innovations are concentration-decreasing than concentration-increasing, the net effect of all of them may be to increase concentration, if the concentration-increasing innovations are more important in the relevant industry and time period.

I looked in some detail at the sources of the drug innovations in our sample. In over half of the cases, the innovators were established firms entering markets that were new to them. In another one-sixth of the cases, the innovator was in the relevant market, but not among the top four firms in that market. The large proportion of cases where the innovator was a new entrant to the relevant market or a relatively small seller in that market is, of course, one reason for the drug industry's relatively low proportion of product innovations that increased the four-firm concentration ratio.

### III. The Rate of Technological Change, the Character of Process and Product Innovation, and Changes in Concentration

According to some economic models, concentration levels are more likely to increase in industries and time periods characterized by relatively rapid technological advance than in those characterized by relatively slow advance. If this is true, one might expect that the proportion of product innovations that are concentration-increasing (and perhaps the proportion of process innovations that are scale-increasing) would be higher in industries and time periods where technological change is rapid than in those where it is slow. To find out whether this was the case in the industries considered here, I took various periods (each about 10–15 years long) in each industry for which data are available concerning the rate of increase of total factor productivity. Then it was determined whether the rate of productivity increase in a period is related to the percent of new products that were concentration-increasing during this period. (Also, I determined whether it is related to the percent of new processes that were scale-increasing during this period.) If the rate of productivity increase is a reasonable measure of the rate of technological change in these industries, this analysis should provide some of the first direct evidence on this score. However, in view of the small number of industries included, the results should obviously be treated with the utmost caution.

It turns out that there is essentially no correlation in these industries between the rate of productivity growth, on the one hand, and the percent of product innovations that were concentration increasing (or the percent of process innovations that were scale increasing). Since the rate of productivity increase may not be a very good measure of the rate of technological change, I used the ratio of *R&D* expenditures to sales instead. The results are much the same. There is no significant relationship between this ratio and the percent of product innovations that are concentration increasing (or the percent of process innovations that are scale increasing). Indeed, there is a negative (but statistically nonsignificant) relationship between the ratio of *R&D* expenditures to sales and the percentage of product innovations that were concentration-increasing. Moreover, if the time periods used in the analysis are lengthened to about twenty years, the results are essentially the same.

Turning to all 2-digit manufacturing industries, is there in fact a close relationship between an industry's rate of technological change (as measured by its rate of productivity increase and its ratio of *R&D* expenditures to value-added) and the change in its average four-firm concentration ratio? To find out, I considered various periods for which data are available concerning the average annual rate of increase of total factor productivity. The coefficient of correlation between an industry's rate of increase of productivity and the change in its average four-firm concentration ratio turns out to be generally negative and never positive. If an industry's ratio of *R&D* expenditures to value-added is used (in place of the rate of productivity increase) as a measure of the rate of technological change, the results are the same. The coefficient of correlation between this ratio and the change in an industry's average four-firm concentration ratio is negative and far from statistically significant. Whether these results would change appreciably (and if so, how) if the effects of other variables were held constant is an open question (and an important one) that lies outside the scope of this paper.

To prevent misunderstanding, it is important to recognize that these results do not deny that an increased rate of technological change is often associated with increased concentration. Without question, such an association often exists. But whether it exists depends on the nature and sources of the new technology. Unless we know something about these and other variables, prediction of the effects of technological change on concentration is likely to be hazardous. One reason why existing models predict that innovation tends to increase concentration is that they often assume that no entry exists in the market under consideration. (Unfortunately, they assume too that no real product innovation occurs.) As Nelson and Winter point out, "things would clearly be different if entrants came in at large scale, as technological leaders, and motivated by subtle long-run strategic considerations" (p. 543). This, in fact, is what has happened frequently in the drug and chemical industries (as shown in Section II). Moreover, this sort of "innovation by invasion" occurs in many other industries too. In situations of this sort, innovation may (as we have seen) reduce existing concentration levels, not increase them. Difficult though it may be, the interesting models that have been constructed in recent years should be extended to take these important factors into account.

In conclusion, the empirical results presented in this paper should be regarded as tentative first steps; much more should be done. However, I doubt very much that future work will alter the principal point of this discussion, which is that, unless we know the nature and sources of new technology, the prediction of the effects of technological

change on concentration is hazardous indeed.

## REFERENCES

- Blair, John, *Economic Concentration: Structure, Behavior, and Public Policy*, New York: Harcourt Brace Jovanovich, 1972.
- Burns, Arthur, *The Decline of Competition*, New York: McGraw-Hill, 1936.
- Galbraith, John Kenneth, *The New Industrial State*, Boston: Houghton Mifflin, 1967.
- Landau, Ralph, "Chemical Industry Research and Development," in W. N. Smith and C. Larson, eds., *Innovation and U.S. Research*, Washington: American Chemical Society, 1980.
- Levin, Richard, "Toward an Empirical Model of Schumpeterian Competition," Yale University, March 1980.
- Mansfield, Edwin, *Industrial Research and Technological Innovation*, New York: W. W. Norton, 1968.
- \_\_\_\_\_, et al., *Research and Innovation in the Modern Corporation*, New York: W. W. Norton, 1971.
- \_\_\_\_\_, et al., *The Production and Application of New Industrial Technology*, New York: W. W. Norton, 1977.
- \_\_\_\_\_, et al., *Technology Transfer, Productivity, and Economic Policy*, New York, W. W. Norton, 1982.
- Nelson, Richard and Winter, Sidney, "Forces Generating and Limiting Concentration under Schumpeterian Competition," *Bell Journal of Economics*, Autumn 1978, 9, 524-48.
- Schwartzman, David, *Innovation in the Pharmaceutical Industry*, Baltimore: Johns Hopkins, 1976.

# *R&D and Productivity Growth: Policy Studies and Issues*

By ROLF PIEKARZ\*

In the past twenty-five years, the topic of the role of research and development (*R&D*) in productivity growth has emerged from near obscurity to a prominent public policy concern and a substantial research enterprise. Today, many government actions are advocated to increase *R&D*; frequently, these actions are justified on the basis of the contribution of *R&D* to more rapid productivity growth. Also, there is substantial research activity in economics on two classes of conceptual and empirical questions: the effects of *R&D* on productivity growth; and the influence of various market conditions, economic organizational structures, and government policy instruments on *R&D* activity and outputs. Few of the individuals involved in these policy and research processes have taken time to consider explicitly how the accumulating research findings affect the public policy debates on *R&D* and productivity growth.

In this brief paper, I will present two examples to illustrate three ways economic research on questions relating to technological change interacts with policy deliberations. First, advocates of policy positions appeal to research findings in presenting their positions. Second, officials look to research to help project and assess probable outcomes of policy actions in order to improve upon their decisions. Third, both researchers and government officials seek to learn from experience with policy changes in order to reduce underlying uncertainties. My discussion will touch only a few highlights; they can be taken as examples of the application of social science research to the public policy process.

\*Senior Staff Associate, National Science Foundation. Any opinions, findings, conclusions or recommendations expressed in this paper are my own, and do not reflect the views of the National Science Foundation. A complete bibliography may be obtained from the author upon request.

## I. Background

In the face of inflation and increasing unemployment, a great deal of attention has been given during this past decade to the slowdown in productivity growth in the United States. Public policy discussion has placed great weight on *R&D* as a means to accelerate productivity growth. In these discussions, the results of research on *R&D* and productivity growth during the 1960's by economists, such as Zvi Griliches, Edwin Mansfield, and Jora Minasian, helped importantly to shape perceptions of the role of *R&D*. Another influential line of research was the work of Edward Denison, John Kendrick, and Dale Jorgenson on the measurement of the elements of U.S. output growth, and factors contributing to that growth.

One element of the discussion of the role of *R&D* in productivity growth is the existence of two different views about what aspect of *R&D* activity has been most important to the productivity growth slowdown. Some experts have focused on a slowdown in growth, or even reduction, in "real" U.S. *R&D* spending during the 1970's. Others have claimed that the United States, like Great Britain, has not adequately exploited the economic potential of its *R&D* capabilities and *R&D* results.

This public debate has resulted in greater government attention to enacting and implementing policies both to stimulate *R&D* and to increase the productivity benefits from *R&D*. Examples of policies advocated to stimulate *R&D* include: incentives provided by changes in tax treatment of business funding of *R&D*; and legal measures to eliminate certain prohibitions on firms conducting *R&D* (for example, allow large firms to engage in joint *R&D* ventures) or on firms' capture of the economic benefits from *R&D* (for example, less economic regulation of the communications and transportation sectors).

Examples of policies advocated to increase the productivity from *R&D* include federal programs to stimulate the diffusion of findings from government funded *R&D* (for example, industry-university research centers); and federal incentives to states and firms to upgrade the scientific and technical skills of the labor force.

Two recent government policy thrusts form the examples discussed in this paper. One initiative is the change in the corporate tax treatment of *R&D* in order to stimulate business *R&D*. The other consists of some recent institutional changes initiated by the federal government in order to increase the contribution of government *R&D* to productivity growth.

## II. Corporate Income Tax Treatment of *R&D*

Prior to the enactment of the Economic Recovery Tax Act of 1981 (ERTA), some critics argued that the tax code did not do enough to mitigate corporations' underinvesting in *R&D*. No one maintained that *R&D* received unfavorable tax treatment. Under both pre-ERTA and ERTA law, current outlays for *R&D* (for example, wage and salary payments) can be deducted in the year incurred the same as current expenses for production and marketing.

But, proponents of more favorable tax treatment for *R&D* held that the tax system failed to mitigate the disparity between private profitability and social returns from *R&D*. As a result, the tax system encouraged corporations to allocate funds to activities having lower social returns than *R&D*. This side of the debate maintained that relative tax treatment of various types of earnings and expenditures importantly influences a firm's decisions. With respect to the divergence between private profitability and social rates of return, advocates of more favorable tax treatment mentioned the findings by Henry Grabowski which suggested that spending on advertising resulted in about the same profits to firms as spending on *R&D*. Advocates also cited findings by Mansfield which suggested a low probability of commercial success associated with *R&D* projects by firms. Social returns were pre-

sumed higher for *R&D* outlays. Frequently cited to support this claim were the findings of the studies by Mansfield, and the follow-on efforts by John Beyer and James Tewksburg, which showed that *R&D* expenditures by firms usually resulted in social returns substantially in excess of the profitability to firms performing the *R&D*. These findings were consistent with the earlier work of Mansfield, Minasian, and Nestor Terleckyj which showed substantial increases in outputs in manufacturing from business-funded *R&D*.

Proponents of more favorable tax treatment for *R&D* then pushed their argument a step further contending that specific tax incentives for *R&D* would encourage firms to increase *R&D* spending, and thereby provide greater social returns from their activities. Proponents tended to focus on questions about what form the tax incentives should take rather than on whether and in what circumstances a tax incentive might be expected to elicit additional *R&D* activity. For example, one question debated heatedly was whether all business *R&D* expenditures or only a certain subset (for example, increased *R&D* spending) should be given more favorable tax treatment.

Of course, not everyone agreed that specific tax incentives for *R&D* would be a cost-effective approach to stimulate private sector *R&D* activity. Opponents of specific incentives based their argument on several lines of analysis. A microeconomic approach was used to argue that *R&D* activities represented too small a portion of the total costs of innovation projects for tax treatment of *R&D* to make much difference to firms' decisions about technological innovation. Frequently cited were studies by Mansfield and Robert Charpie et al., which indicated that *R&D* expenditures tended to be only a fraction of the costs required for the commercialization of new or improved products or processes. Complementary views about the critical importance of costs of introducing and implementing the results of *R&D* were drawn from a large number of empirical studies of specific technologies. Some prominent names include Bela Gold (steel-making processes) and Mansfield (machine tools).

Opponents of specific incentives also suggested that market opportunities in terms of growing demand might dominate firms' decisions about technical change, including *R&D*. Here, the ideas of Jacob Schmookler were used frequently.

The question of whether and in what circumstances tax policy might influence *R&D* was addressed in two studies that were widely consulted by groups drafting positions in the pre-ERTA debates about appropriate tax treatment for business *R&D*. Both studies, one by Robert Kaplan et al., and the other by the National Academy of Engineering, assessed what available research evidence could say about the influence of tax policy on *R&D* and on technical innovation. These studies concluded that there was no evidence from available research results which supported a strong conclusion about the impact of *R&D*-specific tax incentives on *R&D* spending. These surveys did suggest that enabling factors (for example, growing market demand) could importantly influence business *R&D*, and even possibly the impact of *R&D* tax treatment on firms' *R&D* decisions.

Results of this work were used with results of some research on investment behavior to suggest a second approach to stimulating *R&D*—more favorable tax treatment of plant and equipment. Proponents of this approach maintained that much productivity-enhancing *R&D* is incorporated in new plant and equipment; as a result, tax measures favoring plant and equipment investment (for example, investment tax credits, accelerated depreciation) would expand the market for capital goods and thus encourage increased *R&D* in the capital goods sector and speed the rate of adoption of new technology. Research frequently consulted here has been the work of Schmookler and various versions of the empirical analyses pioneered by Jorgenson.

In the light of these uncertainties about the effects of specific tax incentives for *R&D*, more favorable tax treatment of business *R&D* expenditures was included in ERTA only on an experimental basis and only in the later stages of the congressional legislative process. With regard to business tax-

ation, ERTA focuses primarily on tax treatment of plant and equipment investments. For business *R&D* spending, the major provision is a 25 percent tax credit for increases in eligible *R&D* expenditures. The 25 percent tax credit has a "sunset" date; it lapses after 1985. There are three other provisions relating to tax treatment of *R&D* spending. ERTA provides an increased charitable deduction for donation of new equipment by the manufacturer to an institution of higher education for use in research or research training in the natural sciences. Also, the Act allows firms to depreciate equipment used for *R&D* in three years. Finally, the Act provides that, for two years, domestic *R&D* expenditures are not subject to Treasury Regulation 1.861-8. This regulation requires U.S. multinational corporations to apportion part of domestic *R&D* expenditures to foreign-source (rather than domestic-source) income for purposes of computing the foreign tax credit. Since this can reduce a corporation's foreign tax credit, but not its foreign tax liability, it can increase the corporation's total tax liability.

Enactment of the investment and *R&D* incentives has provided a remarkable research opportunity. First, we can use the ERTA experience for systematic analyses of some of the major questions unanswered in the debate prior to ERTA and for retrospective policy assessment. Second, there is sufficient time for fundamental research to provide answers which can usefully be applied in the deliberations about "861" and about the renewal of the *R&D* tax credit. Examples of some important questions for research are as follows: 1) How much additional *R&D* is business likely to fund per dollar of revenue loss from the 25 percent tax credit? 2) Under what circumstances will firm *R&D* expenditures be responsive to the *R&D* tax credit? 3) Do the tax provisions relating to plant and equipment investment "favor" investment in the traditional capital intensive industries over investment in the dynamic *R&D* intensive industries? 4) In what circumstances, if any, do the ERTA investment incentives tend to offset rather than reinforce the *R&D* incentives?

### III. Commercialization of Government-Funded R&D

From the 1950's through the mid-1970's, the federal government was the primary source of R&D funds; only in recent years, has business supplanted the federal government as the major source of funds. With the slowdown of U.S. productivity growth in the 1970's, attention began to be directed to the negligible measured contribution to productivity of most large-scale federal scientific and technological programs. Attempts by economists such as Lawrence Goldberg, Griliches, and Terleckyj to obtain a relationship between government-funded R&D and productivity in industry were not successful in finding such a relationship.

A number of studies have appeared in the past decade questioning the economic effectiveness of government funding of technological development, especially large-scale projects intended for commercial application. Early in the 1970's, papers by George Eads, Stephen Enke, and Richard Nelson indicated that government funding was not warranted for large-scale technological development projects with commercial application. Case studies of large scale government funding of commercial technological development in the United States (Peter House, Nelson et al.), France (John Zysman), and Great Britain (Keith Pavitt) found a poor payoff from large-scale government involvement.

Until the past three to four years, the findings of these studies had little effect with regard to the federal government undertaking funding development of large-scale commercial technologies. In the early 1970's, the federal government undertook a number of large-scale programs for technological development in housing and mass transportation. Later, these efforts were succeeded by large-scale programs to develop nuclear, solar, and synthetic fuels technologies. Though the findings of studies of economists were ignored in these instances, the ultimate outcomes of these programs tend to reinforce the correctness of the economists' results.

In contrast, there are studies which indicate that federal government-funded R&D can contribute importantly to commercial technology and productivity growth under certain circumstances. Recent research is deepening and broadening these findings. Economists, such as Robert Evenson, Vernon Ruttan, and Griliches have demonstrated a relatively high economic return from government-funded R&D in agriculture. During the 1960's and 1970's, the NSF sponsored a number of case studies (for example, TRACES I and II) which showed the substantial contribution of basic disciplinary research to technological development. Recently, Richard Levin, Albert Link, Mansfield, and Terleckyj have published findings of empirical analyses showing that government-funded R&D contributes to development of commercial technology by firms. In parallel, Nelson et al. published a number of case studies showing the same findings. In these various studies, factors influencing the magnitude and/or the timeliness of the commercial exploitation of the results of the federal R&D activity have been hypothesized but not subjected to comprehensive empirical analysis.

These results, coupled with findings of other studies, influenced some recent tentative federal government efforts at institutional change in the R&D system. Studies about the smallness, or the lag, in commercial application of federally funded R&D performed in industry and in universities contributed to stimulating efforts to improve the dissemination of these scientific and technological developments. Important models in proposing measures were the institutions to disseminate the scientific and technological results from agricultural R&D to practice. Such institutions did not exist elsewhere. One recent attempt to accelerate the flow of scientific findings to industry has been the National Science Foundation program to set up joint industry-university centers in a domain of applied science or technology (for example, polymers, robotics) to accelerate the translation of fundamental scientific research to industry. The Stevenson-Wydler Act in 1980 sought to substan-

tially expand this experimental program throughout the federal government, but budget stringency has kept this effort on a small scale and confined to the NSF. The NSF program staff is currently assessing the cost effectiveness of these various efforts.

Another initiative was the Bayh-Dole Act of 1980, which permitted small businesses and nonprofit organizations (including universities) to retain title to patents resulting from federally supported research and to grant limited exclusive licenses. Prior to the Act, policies with respect to patent title varied across agencies, and there were concerns about the poor record of commercial exploitation of federally held patents. Some observers claimed that ambiguity about exclusivity of patent rights discouraged commercial use of worthwhile inventions. The effect of this Act on the commercial application of government-funded *R&D* is not being monitored.

Further progress in devising institutional structures or devices to improve the commercial gains from government *R&D*, must await a better understanding of the economic and legal factors influencing business *R&D* decisions. To date, economists have restricted their empirical work here to fairly straightforward questions about the interaction between firm size or antitrust regulation on *R&D* activity. Most notable is the work of F. M. Scherer and Mansfield. During recent years there has been a good deal of conceptual work in economics about the decision processes of firms with regard to *R&D* and

technological innovation. Prominent in this line of research are Nelson, Sidney Winter, and Joseph Stiglitz. But little of this conceptual work has as yet been translated into empirical research. We are now experiencing a substantial surge of such empirical work and trust that in the latter years of this decade the findings of this research will contribute to policy debate and to options for government policies to enhance the contribution of *R&D* to productivity growth.

#### IV. Concluding Observations

I would like to conclude with a couple of observations based on my experiences with the policy studies and issues sketched in this paper. First, the primary contribution of research is to inform the participants in the policy process in their debates and deliberations. In particular, studies on questions arising in policy discussions often improve the framing of the questions and issues under consideration. Policy studies rarely provide conclusive resolution of a policy problem.

Second, the research community should play an active role in formulating the priority research questions relevant to a policy issue. The protagonists tend to have an incomplete understanding of the important relationships and the state of the relevant scientific knowledge. The usefulness of research to policy deliberations is greatly enhanced when the interested researchers are familiar with the issues under discussion.

# *R&D and Declining Productivity Growth*

By F. M. SCHERER\*

This paper attempts to assemble some pieces of a puzzle. The broad pattern is well known. The growth rate of private business sector labor productivity, which averaged 3.1 percent per annum between 1947 and 1968, fell in stages since then and in 1977–81 barely exceeded zero. Concurrently, real (i.e., *GNP* deflator-adjusted) company-financed industrial research and development (*R&D*) abruptly stopped growing in 1970 and, after a two-year hiatus, resumed its growth at a much slower rate, at least until recently. Had the 6.3 percent annual *R&D* growth rate of 1960–69 persisted during the 1970's, U.S. industry in 1981 would have performed 38 percent more real *R&D* than it actually carried out.

Economists have long been persuaded that *R&D* yields technological advances that in turn foster productivity growth. But the magnitudes involved have been poorly understood, so it has been difficult to say how much of the recent productivity growth slump has resulted from the *R&D* slowdown. Also, why did *R&D* growth stagnate? Could it have been because *R&D* lost some of its power to propel productivity growth? And if so, what are the productivity ramifications?

In this paper I summarize the implications of my own effort to shed light on these questions. Separate research thrusts were directed toward the causes of the *R&D* slowdown and toward *R&D* productivity links. The details are published elsewhere; here integration and interpretation are emphasized.

## **I. The Profitability of *R&D***

Company-financed *R&D* is without doubt a profit-seeking activity. Using pooled time-series–cross-section PIMS data, David Rav-

enscraft and I (1982) attempted to measure as precisely as possible the lag structure of *R&D*—profit links and the biases imparted by applying an improperly specified lag structure. The best-fitting lag was binomial in shape, with peak profits accruing four to six years after *R&D* spending. For a cross section of business units with the longest time-series, we found pretax internal rates of return on *R&D* to be negative for profits realized in 1975, increasing into the 27–45 percent range by 1977–78. The increases in estimated profitability are statistically as well as economically significant.

For the early 1970's, the time-series are much less complete. However, results with naive same-year lag structures suggest that in at least 1970 and 1971, the returns to *R&D* were, as in 1975, severely depressed. This evidence is consistent with sample business units' behavior: they cut back their *R&D* spending relative to sales growth. Through the elimination of low-yield projects, the profitability of *R&D* then rose in the decade's second quinquennium, inducing (with a lag) spending increases that became evident in 1979. From then through 1982, the real growth rate has been approximately 5.7 percent per year.

## **II. *R&D* and Productivity Growth**

Ascertaining how *R&D* affects productivity growth poses more difficult conceptual and data problems. One problem stems from the fact that roughly three-fourths of all company-financed industrial *R&D* is oriented toward the creation of new and improved products which are then sold to other industries. Because competition erodes innovators' rents and because the price deflators used to prepare productivity indices often fail to capture the superiority value of improved products, a considerable fraction of the productivity benefits from an industry's product *R&D* is likely to be cap-

\*Professor of economics, Swarthmore College. The research underlying this paper was supported by NSF grant PRA-7826525.

tured by other industries purchasing its products. Relating  $R\&D$  in industry  $i$  to industry  $i$ 's own productivity growth will therefore yield downward-biased impact estimates. To measure the productivity benefits more comprehensively, a matrix tracing 1974  $R\&D$  from industries of origin to industries of use was estimated (see my 1981 paper). Identifying such interindustry technology flows is particularly important in studying productivity growth for nonmanufacturing industries, which do little  $R\&D$  on their own, but "import" new capital goods or materials embodying roughly half of the manufacturing sector's  $R\&D$ . Within manufacturing, approximately 58 percent of used  $R\&D$  consists of internal process  $R\&D$ ; the rest is imported from other industries. Even for process  $R\&D$ , however, industry-wide productivity benefits may exceed private profit returns because of licensing and/or competitive imitation that lead to reduced end product prices.

Once  $R\&D$  had been traced to industries of use, a productivity regression could be estimated, following Nestor Terleckyj (1974), in the basic form:

$$\Delta LP = \lambda + \frac{\partial Q}{\partial RS} \frac{RE}{Q} + \alpha \Delta(K/L),$$

where  $\Delta LP$  is the annual percentage change in output per work hour,  $\Delta(K/L)$  is the change in the real capital/labor ratio,  $RE/Q$  is the ratio of  $R\&D$  flows to output value, and  $\partial Q/\partial RS$  is the marginal productivity of the  $R\&D$  capital stock (measured as a rate of return on  $R\&D$  capital). The  $\partial Q/\partial RS$  term is of prime interest and was estimated for product  $R\&D$ , process  $R\&D$ , and  $R\&D$  "imported" from other industries using two different disaggregated productivity data sets.

The results are reported fully in my 1982 article; here only highlights can be summarized. A "wrong lag" specification tested the hypothesis that the marginal productivity of  $R\&D$  declined from the 1960's to the 1970's. The hypothesis was not supported;  $\partial Q/\partial RS$  estimates for 1973–78 productivity growth were usually higher than, or in one data subset, only inappreciably less than, the corresponding estimates for 1964–69. Esti-

mated 1973–78 returns on product  $R\&D$  were close to, and insignificantly different from, zero except in one productivity data subset of doubtful reliability. Returns on imported  $R\&D$  were generally higher than on own-process  $R\&D$ , but the differences were not statistically significant. Combining the two into a "used"  $R\&D$  measure, the estimated  $\partial Q/\partial RS$  values imply 1973–78 social returns on used  $R\&D$  in the range of 71 to 104 percent, except for the productivity data subset of most dubious reliability. Thus, the analysis points toward high marginal rates of return to used  $R\&D$  during the 1970's, and, relative to both the estimated private profitability of  $R\&D$  and the productivity index-based returns to product  $R\&D$ , a large divergence between social and private returns.

### III. Interpretation

I advance now to new ground. The estimated  $\partial Q/\partial RS$  coefficients can be used to approximate the  $R\&D$  growth slowdown's impact on productivity. Had  $R\&D$  continued to grow at its 1960's trend rate, 1974  $R\&D$ /output ratios would have been 1.37 times their actually observed values. Mean *used*  $R\&D$  in 1974 as a percentage of sales for my most comprehensive industry sample (covering nearly all of manufacturing plus agriculture, crude oil and gas, railroads, airlines, telecommunications, and electric-gas-sanitary utilities) was 0.73. Multiplying  $0.73 \times (1.37 - 1)$ , I estimate the 1974  $R\&D$  shortfall to have been 0.27 percentage points. For that sample, the used  $R\&D$   $\partial Q/\partial RS$  coefficients ranged between 0.74 and 1.04. Thus, the 1978 productivity growth shortfall attributable to lower  $R\&D$  is estimated to be on the order of 0.20 to 0.28 percentage points per year. Or if the coefficients of a smaller industry sample with particularly well-measured productivity indices (*BLSPQ* in my 1982 article) are used, the  $R\&D$ -related productivity growth impairment is estimated to be about 0.39 percentage points. Since the ratio of 1960's trend-extrapolated  $R\&D$  to actual  $R\&D$  in the later 1970's continued to lie in the range of 1.37 to 1.42, a drag on annual productivity growth of similar magni-

tude (i.e., in the range of 0.2 to 0.4 percentage points) can be expected to continue for some time.

These estimates assume *inter alia* that the coefficient  $\partial Q/\partial RS$  estimates the *marginal* productivity of R&D (rather than, say, the average return) and that the marginal productivity of R&D remains constant over a rather large change in the amount of R&D performed and utilized. This assumption may be wrong, but the bias from the error is not as obvious as one might suppose. It depends upon the particular mechanism inducing a slowdown of R&D growth. To illustrate, two alternative scenarios will be examined.

A central feature in any case must be a fall in the private profitability of R&D. We must also recognize that there is a divergence between marginal social and private rates of return. Based upon calculations using the data reported by Edwin Mansfield et al. (1977, p. 233), it is assumed that private returns deviate from social returns by a constant fraction, so that the private marginal efficiency of R&D investment schedule *MPER* intersects the marginal social efficiency schedule *MSER* at a zero rate of return, with larger absolute departures at higher rates of return. This assumption is also roughly consistent with dynamic limit pricing theory, assuming entry barriers (for example, patent protection) to be uncorrelated with social rates of return.

In Figure 1, the initial marginal efficiency schedules are *MSER*<sub>1</sub> and *MPER*<sub>1</sub>. With a constant real risk-adjusted cost of capital *Oi*, the profit-maximizing amount of R&D investment is *OR*<sub>1</sub><sup>\*</sup>. If the event precipitating a fall in R&D spending is a parallel leftward shift in *MSER*, for example, because the pool of innovation opportunities became depleted or because product characteristics space became more densely packed, the new marginal efficiency schedules will be *MSER*<sub>2</sub> and *MPER*<sub>2</sub>. The profit-maximizing level of R&D will fall to *OR*<sub>2</sub><sup>\*</sup>, and the net social surplus loss will be the shaded trapezoidal area *ZP*<sub>1</sub>*BA*. Because of the assumed parallel marginal efficiency schedule shifts, the marginal social return on R&D after the market has adjusted will be observed at level *R*<sub>2</sub><sup>\*</sup>*P*<sub>2</sub>

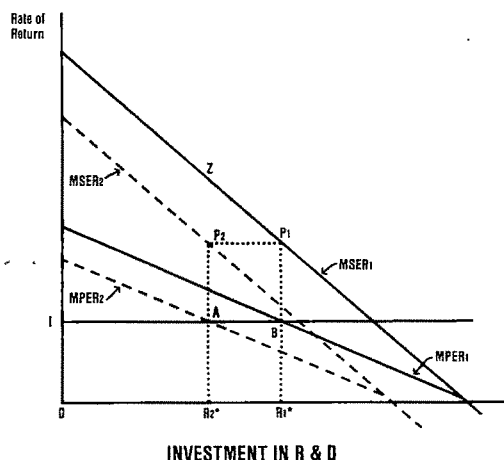


FIGURE 1

—identical to the preshift estimate *R*<sub>1</sub><sup>\*</sup>*P*<sub>1</sub>. But relative to the earlier, rich-opportunity situation, a larger social loss *R*<sub>2</sub><sup>\*</sup>*Z* will be incurred at the margin. Using the observed *ex post* marginal social return on R&D to measure the economywide impact, as I have done above, will lead to an underestimate by the amount of the triangle *ZP*<sub>1</sub>*P*<sub>2</sub>.

Figure 2 provides an alternate scenario in which R&D retains its social productivity and the impetus to declining R&D support is an increasing divergence between private and social returns—for example, because of intensified research competition (*inter alia* from abroad) or more rapid imitation. The social loss is again measured by trapezoidal area *P*<sub>2</sub>*P*<sub>1</sub>*BA*. But now the postshift marginal social return *R*<sub>2</sub><sup>\*</sup>*P*<sub>2</sub> will *overestimate* the social loss on foregone R&D projects positioned between *R*<sub>2</sub><sup>\*</sup> and *R*<sub>1</sub><sup>\*</sup>.

Evidently, ascertaining more precisely how the R&D slump has affected productivity and economic welfare requires a better understanding of *why* R&D spending stagnated. And more importantly, the choice of appropriate corrective policies—for example, increasing support of basic research vs. strengthening patent protection—demands such insights. The increases in estimated  $\partial Q/\partial RS$  coefficients between 1964–69 and 1973–78 lend support to a Figure 2 increasing divergence hypothesis, but the data are

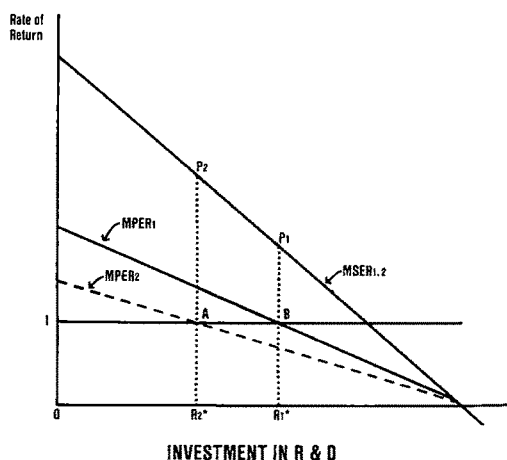


FIGURE 2

not strong enough to warrant a conclusive inference. The profitability study by Ravenscraft and myself yielded hints that private *R&D* returns fell in part because of more intense competitive pressure, but they were no more than hints. Qualitative observation and some quantitative research (for example, on pharmaceuticals) suggests that in at least some fields of technology, there has in fact been a depletion of opportunities. See my 1978 article. Nevertheless, much more work is needed before we can have an adequate picture of why *R&D* spending growth flagged and what its exact connections are to the productivity hammer-blow with which it is undoubtedly linked.

If we are to make significant progress on that front, we must have better data. The best disaggregated data on *R&D* spending come from the Federal Trade Commission's Line of Business program, covering the years 1974-78. The program's continuation is now in jeopardy. The Canadian Patent Office has since 1978 been collecting comprehensive in-

vention patent origin and use data like those I used to estimate my technology flows matrix. In the United States, no similar data development effort exists. And most important of all, our industry productivity series leave much to be desired. I found quite different *R&D*-productivity growth link patterns in industry subsets whose productivity indices were based upon reasonably comprehensive product price deflators, as compared to those with sparse deflators. While Rome burns, we do little to assemble better information on important fire extinguishing materials. Perhaps the fire will burn itself out, but in the meantime, appalling losses accrue.

## REFERENCES

- Mansfield, Edwin et al., "Social and Private Rates of Return from Industrial Innovations," *Quarterly Journal of Economics*, May 1977, 91, 221-40.
- Ravenscraft, David, and Scherer, F. M., "The Lag Structure of Returns to *R&D*," *Applied Economics*, December 1982, 14, 603-20.
- Scherer, F. M., "Technological Maturity and Waning Economic Growth," *Arts & Sciences*, Northwestern University, Fall 1978, 7-11.
- \_\_\_\_\_, "Using Linked Patent and *R&D* Data to Measure Inter-Industry Technology Flows," paper presented at a National Bureau of Economic Research Conference on *R&D*, Patents, and Productivity, October 1981.
- \_\_\_\_\_, "Inter-Industry Technology Flows and Productivity Growth," *Review of Economics and Statistics*, November 1982, 64, 627-34.
- Terleckyj, Nestor, *Effects of *R&D* on the Productivity Growth of Industries*, Washington: National Planning Association, 1974.

## MACROECONOMICS: MAJOR ISSUES AND DEVELOPMENTS

### Is Unemployment a Macroeconomic Problem?

By ROBERT E. HALL\*

Rather than start directly on the sensitive issue of the economic role of unemployment, I would like to spend some time first on a parallel question of rather less social importance, and then draw some analogies to the problem of unemployment. The phenomenon I will examine is the time people spend idle at airports. Ultimately, I will compare the analysis of idle airport time with the analysis of idle time in the labor market.

In any airport at any time, numerous people are waiting for something to happen. These people are not doing anything particularly constructive with their time—they are waiting because they arrived early, because their planes have been delayed, or because they are in a queue for the next available flight. An observer who knew nothing about the purpose of an airport would be puzzled by the chronic idleness of most of the people there. The observer might gather data on airport idleness along the following lines. At any given time, 0.2 percent of the population is idle at the airport. The idle population turns over frequently—the median duration of a spell at the airport is 35 minutes. But long spells account for the bulk of idleness—half of all idleness occurs in the course of spells which will last 5 hours or more. Airport idleness is highly concentrated in the population. In a given year, three-quarters of the population are never idle at the airport; 5 percent of the population incurs half of all idleness. A predictable seasonal pattern is apparent—idleness reaches sharp peaks at Thanksgiving, Christmas, and Easter, plus a broad peak in the summer.

Were it quantitatively more significant, airport idleness would be a social issue. The airport idle are not usually engaged in useful activities. Few of them spend time trying to locate earlier flights, nor do many of them try to accelerate their movement by offering to pay a higher fare. A surprisingly large fraction do nothing more than sit. The opportunity cost of time spent idle at the airport is essentially zero, it would appear.

#### I. A Microeconomic Model of Airport Idleness

People come to airports because they plan to take a flight, or because they are picking up somebody. The flow to the airport is controlled rather closely by the demand for air travel. That demand can reasonably be taken as a function,  $D(p, r)$ , of air fares,  $p$ , and the utilization rate or load factor,  $r$ . A higher utilization rate repels travelers because it is more difficult when more flights are full to make convenient arrangements, to change plans, or to make up for a missed plane.

The supply of airline seats is also a function,  $S(p, r)$ , of the fare  $p$ , and the utilization rate  $r$ . In the long run, supply may reasonably be taken to be perfectly elastic—if revenue per seat,  $pr$ , equals long-run marginal cost, any number of seats will be supplied. In the shorter run, supply will be less elastic. In either case, there will be a downward-sloping schedule in  $r-p$  space depicting all combinations of  $r$  and  $p$  which equate supply and demand, as shown in Figure 1. At the upper left, the market clears with high fares and low utilization; at the lower right, with low fares and high utilization.

Equilibrium would be indeterminate if airlines were price takers and utilization rate takers in the market. But they are not. If all the other airlines are charging  $p$  and the

\*Hoover Institution and department of economics, Stanford University, and National Bureau of Economic Research.

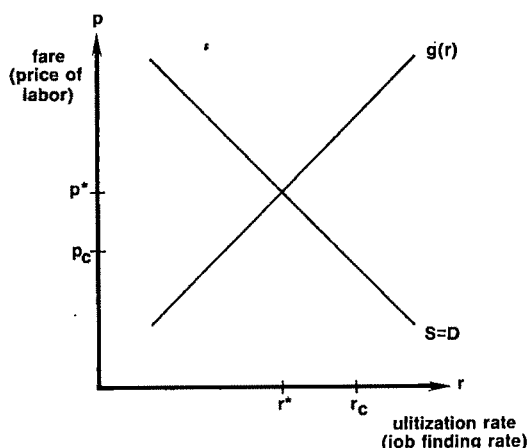


FIGURE 1

market is clearing with a utilization rate of  $r$ , it may be optimal for any given airline to set a fare different from  $p$  and achieve a utilization rate different from  $r$ . If prevailing fares are high and industry utilization is low, it may be attractive to set a lower fare and fill up seats. If fares are low and utilization is high, passengers will flock to an airline that has seats readily available at somewhat above prevailing fares. As a general matter, one airline ought to set its fare according to a function  $\phi(p, r)$  of the fare set by all others,  $p$ , and the state of the market as measured by  $r$ . This function is increasing in  $p$ , but with derivative less than unity. It is increasing in  $r$ . In equilibrium, all airlines set the same fare, which means  $p = \phi(p, r)$ . We can solve this equation to obtain an upward-sloping schedule  $g(r)$ , also shown in Figure 1. At each point on this schedule, the profit-maximizing fares chosen by each airline, given the fares of others, and the prevailing utilization rate will be equal.

The intersection of the downward-sloping market-clearing schedule and the upward-sloping  $g(r)$  schedule is the unique equilibrium in the market for airline seats. At fare  $p^*$  and utilization rate  $r^*$ , passengers are arriving at the airport at the same rate that airlines are serving them. Further, no airline can improve its profit by setting a fare different from the prevailing fare. Utilization may be well under 100 percent—were it that high,

many prospective passengers would be willing to pay higher fares in order to get seats at the last minute.

The equilibrium depicted in Figure 1 is robust, in the sense that it occurs at the intersection of upward- and downward-sloping schedules. In particular, modest shifts in either schedule do not bring large shifts in utilization rates. Were utilization to rise sharply, it would be a signal of a strong outside impulse, not a move generated internally by the market because of a nearly unstable equilibrium.

The equilibrium in Figure 1 assumes complete freedom on the part of airlines in setting fares to maximize profits. But the figure can also describe the outcome of fare controls. If the regulators prescribe a fare above the free-market level, the market will clear in a more limited sense at a point up and to the left along the  $S = D$  schedule; utilization will be below its optimum. Similarly, the regulators can push the market down and to the right along  $S = D$ , in which case utilization will be too high.

What about airport idleness? Given the prevailing fare  $p$  and the utilization rate  $r$ , at any point along the  $S = D$  schedule (not just the full equilibrium,  $p^*$  and  $r^*$ ), the number of people waiting at the airport is given by the decisions of the public about how much extra time to allow between departure for the airport and flight time. For a number of reasons, the time is higher when utilization is higher. Check-in time is longer when utilization is high, for example. The optimal safety margin is higher when flights are crowded because the consequences of a missed plane are more serious when it will be difficult to line up an alternative flight. Further, airport idleness will rise in times of high utilization precisely because the wait for an available flight is longer after missing a plane.

## II. Is Airport Idleness a Social Problem?

It strikes me as fair to say that the idleness at the full equilibrium in Figure 1 is not a social problem even though it probably means hundreds of millions of man- and woman-hours per years of almost completely wasted time. The idleness is a problem, in

the sense that it would be nice if air travel could be accomplished without idleness, but it is not a situation that calls for relief through intervention.

The more significant question is: how would we evaluate a substantial increase in airport idleness that persisted for a year or more? The analysis makes clear that the answer depends entirely on the source of the change. Nothing in the model tells us that every rise in idleness is a pure waste of people's time. For example, an exogenous drop in airline capacity (say from the airport controllers' strike) would shift the  $S = D$  schedule up and to the right. The new equilibrium would involve higher fares and higher utilization, and so more airport idleness (at least more idleness per passenger mile). In this case, the market is making the best of a bad situation. Though it would be true that idleness could be depressed by putting an emergency tax on air travel, such a move would not be efficient.

On the other hand, it is clear that other forces could bring an increase in idleness that would be a social problem. Suppose that price controls were reinstated in the airline industry, and this time they depressed fares below equilibrium. The industry would operate at a point down and to the right along  $S = D$ , say at fare  $p_c$  and utilization rate  $r_c$ . Airport idleness would jump and now would represent a social problem, analogous to the social problem of long lines at gas stations in 1974 and 1979. The corrective policy needed is obvious—remove price controls and let prices rise and utilization fall.

Now suppose that demand increased, say through general inflation, but that airlines perversely kept their old fares. Then utilization would rise more than would be efficient, and excess airport idleness would follow. Again, higher idleness would be a social problem, but now it is much less obvious what is the appropriate corrective action. The imposition of price controls to force fares upward would do the job in theory, but few economists would trust the regulators to make such a deft intervention after decades of experience with controlled fares far above equilibrium. Economists still lack a good prescription for treating free markets that

fail to do what they are supposed to. About the best we can say about sticky airline fares is that it would be desirable to keep the macroeconomic background as stable as possible.

### III. Unemployment

I hope the analogy between the airport and the labor market is reasonably obvious. The airport idle spend only a small fraction of their time checking with airlines for an earlier flight, just as the job seeker spends little time checking with employers about jobs that might start sooner. Both groups have a pretty clear idea about what is going to develop, and perceive that the right strategy is simply to wait.

The airport resembles the job market in handling a huge volume of traffic routinely at all times. Each week, around a million workers find jobs, just as several million passengers each week accomplish their purposes at airports. In both instances, the flows are stable. In the labor market, workers change jobs according to a stable life cycle pattern. Young people try one job after another until they find a good match, which may then last for decades. The flow into the market from life cycle turnover completely dominates the extra flow from job loss during a recession. Just as a theory of the incidence of airport idleness turns out to be a theory of optimal waiting time, a theory of the unemployment rate is largely a theory of the duration of job seeking. Of course, the time scale is completely different, which is why we worry about unemployment and not about airport idleness.

The model of the airport becomes a model of the labor market with a simple relabeling of the variables. Let  $p$  be the wage and  $r$  be the job-finding rate (the weekly probability that a seeker will find work); these take the place of the fare and the utilization rate. Exactly as before, there is a downward-sloping schedule showing all the alternative combinations of wages and job-finding rates that equate the supply and demand for labor. Higher  $p$  and higher  $r$  are each an attraction for workers and a disadvantage for employers. For an arbitrary wage, the market

will settle on the job-finding rate given by the  $S = D$  schedule.

If the market is at a point up and to the left along the  $S = D$  schedule, an employer can profit by departing from the prevailing terms. A wage below the prevailing wage will still attract workers, because job seekers will face less competition from their comrades when they apply. At points down and to the right, an employer can profit by offering above the prevailing wage—reduced recruiting effort will more than make up for the extra wage cost. Again, there is an upward-sloping schedule,  $g(r)$ , giving the wage level for each  $r$  such that the optimal wage offer for each employer is that wage level. There is a robust equilibrium at the intersection of the two schedules, with job-finding rate  $r^*$  and wage  $p^*$ .

Corresponding to the equilibrium job-finding rate  $r^*$ , is an equilibrium or natural unemployment rate  $u^*$ , which is the product of  $r^*$  and the stable rate of flow into the labor market. Again, only major outside influences can bring a big shift in unemployment. One obvious source of excess unemployment would be the imposition of a binding minimum wage, which would force the market up and to the left along the  $S = D$  schedule. It is hard to think of any other event, comparable to the air controllers' strike, which would cause a year or more of high unemployment.

The United States is about to enter its third year of unemployment far above historical averages, and certainly no government intervention in wages has occurred over

the period to explain it. In terms of Figure 1, the economy is at a point above and to the left of the equilibrium, and the forces that should take us back to the equilibrium are working painfully slowly. Because I cannot think of any forces that might have shifted either  $S = D$  or  $g(r)$  in a way to make them intersect at a much lower  $r$  and higher rate of unemployment, I am forced to conclude that something is wrong—unemployment is a problem.

#### IV. Conclusion

Government intervention in wage setting has such a hopeless history in the United States and elsewhere that I cannot imagine recommending it, even though if by some magic we could coax wages down by 3 percent on January 1, unemployment would fall quickly to a more satisfactory level.

My analysis supports the prevailing consensus of monetarists and Keynesians, that disinflation is a costly process, as against the equilibrium view that the real consequences of price stabilization are transitory and insignificant. Deeply embedded inflationary momentum has pushed us up and to the left along the  $S = D$  schedule; the process that will move us back to equilibrium at a low rate of inflation is a time consuming and costly one. To say in 1982 that unemployment is a major social problem is precisely to say that the decision to inflate the economy in the late 1960's and early 1970's was a costly one.

# Microeconomic Developments and Macroeconomics

By MILTON HARRIS AND BENGT HOLMSTROM\*

We explore the implications of contract theory with regard to the effectiveness of aggregate economic policy. More specifically, we consider the relationship between contractually induced price rigidities and monetary policy. In examining this relationship, we restrict ourselves to *market-clearing equilibrium models*. In particular, we do *not* consider disequilibrium type models such as those of Robert Barro and Herschel Grossman (1976).

Since Keynes' *The General Theory*, economists have blamed unemployment and economic fluctuations on sticky nominal wages. Until recently, however, there were no plausible explanations of this rigidity. About ten years ago, several scholars independently advanced the hypothesis that real wage stickiness might in fact be the result of contractual arrangements. This general idea (in the form of *nominal* wage stickiness) was subsequently applied to show that monetary policy affects real output by inducing additional flexibility in the real wage and can therefore be used to stabilize output fluctuations. While these arguments are based on explicit (rational expectations) macroeconomic models, no comparably articulated model of contracts is provided to account for the nominal wage stickiness. Our purpose here is to examine these arguments critically from the perspective of an explicit contractual model which we have developed elsewhere.

Our main conclusion is that while contractual rigidities may provide a role for activist policy, it is not the role envisioned by the earlier contributors to this literature. In our approach, the role of aggregate policy is to convey information which would other-

wise be more costly to obtain. In particular, the view that monetary policy is effective by directly influencing the real wage is not supported by our model.

## I. A Brief Review of Some Literature

There are two streams of research which will be reviewed here. The first establishes the connection between contractual rigidities and aggregate policy without exploring the reason for the existence of these rigidities. The second focuses more on the microeconomic foundations of contracting.

It has long been claimed that sticky nominal wages could lead to output fluctuations which in turn could be smoothed by appropriate control of the money supply. In the early models, wages are sticky in the sense that they are not contingent on current events, realizations which occur within the period for which the wages are set. Wages can, however, depend on past events. In these models, the money supply (henceforth denoted  $M$ ) can depend on exactly the same set of contingencies as the wage. Even though wages and  $M$  are set based on the same data, the wage-setting sector cannot offset the effects of the monetary policy rule because their price expectations follow some exogenous, adaptive process. Thus, private agents cannot predict the effects of changes in  $M$  on the price level. The monetary policy rule, however, is chosen with full knowledge of how  $M$  affects prices as defined by the model being used to analyze these effects. The result is that "money matters" in those models; the monetary policy rule can be chosen to manipulate the real wage, hence employment, hence output.

Thomas Sargent and Neil Wallace (1975) pointed out that this whole edifice rests on the assumption that agent's expectations are not formed using the same model as the economist (and monetary authority). They adapted an idea of John Muth to show that

\*Professor of finance and managerial economics, and associate professor of managerial economics, respectively, J. L. Kellogg Graduate School of Management, Northwestern University. We are grateful to Robert J. Gordon, Frederic Mishkin, and Laurence Weiss for comments.

when expectations are formed using the same model as the economists, that is, rationally, monetary policy has no effect on employment and output (see also Robert E. Lucas, 1972). In its essence the argument is as follows. If wages and the money supply are determined by the same information, then the money supply for a given period can be perfectly predicted when wages for that period are set. Moreover, under rational expectations, wage setters can predict as well as the monetary authority the effect on the price level of the predicted value of  $M$ . Thus anything done by the monetary authority can be undone by wage setters. Put another way, the monetary authority cannot do anything which wage setters could not have done for themselves given any money supply rule whatever.

This argument makes clear the idea that, even under rational expectations, the money supply rule may matter if the money supply for any given period is not perfectly predictable given the information which determines wages in that period. Edmund Phelps and John Taylor (1977) and Stanley Fischer (1977) exploited this idea to show that money matters. In both models, the wage is assumed to be rigid, not only with respect to concurrently realized events, but also with respect to the information which determines the concurrent money supply. That is, the money supply at any time is based on finer information than the wage at that time. Here we use the term rigid relative to a class of possible realizations (or events) to mean independent of which event in that class actually occurs. For example, in Fischer's model, the nominal wage in period  $t$  depends on shocks realized at  $t-2$  and earlier while the money supply in period  $t$  can depend on shocks realized at  $t-1$  and earlier.

Given the contractual rigidities in these models, the role of monetary policy is simply to act as a substitute for contract indexation. In Fischer's model, for instance, an optimal monetary policy is a perfect substitute for indexing on shocks realized at  $t-1$ , because there is only one relative price that needs to be corrected in each period. With more than one contractual market, however, the single

instrument that monetary policy provides will be insufficient to adjust optimally several relative prices, and a nontrivial tradeoff in policy choice would result.

In none of the models just discussed or cited is there any attempt to model the wage rigidity formally, or to make any connection between the reasons for such rigidity and the manipulation of the environment by the monetary authority. Phelps and Taylor discuss reduction of information costs and price risk as possible reasons. Fischer attributes stickiness to the existence of long-term contracts which result from "costs of frequent price setting and wage negotiations" (p. 194). It has been conjectured that incorporating an explicit model of wage rigidities might significantly alter the conclusions.

Perhaps the first attempt at modeling contractual stickiness in a macro policy framework is that of Joanna Gray (1978). Like Fischer, Gray focused on the length of contracts as the important aspect of contracts leading to wage stickiness. Although Gray does have an explicit model of contract length, the main aggregate policy implications do not flow from this part of the model. The model incorporates the assumption that wages can be indexed to economic disturbances only through adjustments proportional to price level changes. The constant of proportionality (the "degree of indexation") can depend on time, but not on the realizations of the disturbance terms. Since there are two random disturbances in the model, one nominal, one real, this type of indexation cannot insulate the real economy from nominal random disturbances caused by the monetary authority.<sup>1</sup> Moreover, in Gray's model, monetary policy cannot depend on or be correlated with realizations of the real disturbance. The result is that the optimal monetary policy is the least random one (the deterministic aspect of the policy will not matter in this model). Although monetary policy does affect contract length, that is, the period over which the degree of indexation

<sup>1</sup>The inability to index on both nominal and real shocks was rationalized formally using imperfect information by Lucas.

remains fixed, the rigidity which drives the policy implications is exogenous.

What seems to be needed is a more explicit model of why contracts are not indexed on certain information. Ronald Dye (1981) attempted to provide such a model by assuming that contingencies are costly to incorporate. This is shown to have implications both for the length of contracts and the extent to which they depend on realized events. In addition to incorporating explicit contingency costs, Dye also considers the dynamical aspects of the contract length problem (which Gray ignores), namely, that today's contract affects the distribution of possible starting states for the subsequent contract. Unfortunately, the difficulty of pinning down what one means by a contingency and the dynamical aspects make it very difficult to derive comparative statics results or policy implications.

## II. The Rigidity-Policy Interaction

In this section we shall briefly describe a model of contractual stickiness based on work which we present in more detail elsewhere (1982). The basic idea is that contracts are not indexed on certain information because that information is costly to obtain. Such information will be "purchased" (or produced) only occasionally, as opposed to continuously, resulting in contracts of non-zero, finite, and deterministic length. The extent of contractual rigidity is determined by the cost of acquiring information, the extent to which this information can be partially inferred from freely observable events, and the value of the information once obtained. Before pursuing the implications of this approach for aggregate policy, we first sketch the model in more detail.

In its most general incarnation, the model consists of two individuals who must effect some allocation in each of an infinite sequence of periods. The Pareto optimal allocations in each period  $t$  will depend on the current probability,  $\theta_t$ , that a certain event will occur in that period. This probability evolves randomly. Unfortunately, neither of the individuals can observe the probability

directly unless a cost is paid. One of the individuals, the principal, chooses the allocation in each period (subject to a minimum acceptability constraint for the other individual) based only on his current *beliefs* about  $\theta_t$ . He must also choose whether or not to observe  $\theta_t$  (and pay the cost of so doing) in each period.

A multiperiod contract is formed by stringing together a prespecified sequence of allocations for the periods between successive observations of  $\theta_t$ . During a contract, the terms of the contract (the within-period allocation) are rigid with respect to  $\theta_t$ . The current contract ends when either the value of  $\theta_t$  is next observed or when the event whose probability is  $\theta_t$  occurs. This latter event is called a *default*.

To aid our intuition in understanding this model, consider the following example. The principal is a banker and the allocation within each period is a one-period loan (the other individual is a borrower). An optimal one-period loan depends on the borrower's default probability,  $\theta_t$ . The banker cannot discover the borrower's default probability, which changes from time to time in a non-predictable way, without spending resources to perform an up-to-date credit check. He may therefore choose a sequence of one-period loans (a contract) based on his current beliefs about the borrower's credit worthiness and announce that at the end of that sequence, he will once again perform a credit check. These loans cannot depend on the evolution of  $\theta_t$  during the contract. After the next check, subsequent loans will depend on the results of the check. Meanwhile, should the borrower default before the end of the current contract, the specified sequence of loans will terminate immediately and a new contract will begin. This example is only one of many (including labor contracting) that will fit the general paradigm.

The analysis of the model consists in solving the principal's dynamic optimization problem for choosing how often to purchase information. It is shown that under reasonable assumptions on preferences and other parameters of the model (for example, information costs), contracts will last a nonzero,

finite, noncontingent number of periods. This number is shown to depend on the principal's initial information, the extent to which information is revealed by whether or not a default occurs, the parameters governing the random evolution of the default probability  $\theta_t$ , and the cost of obtaining direct information on  $\theta_t$ . Specifically, contract length is shown to increase with increases in direct information costs. Contract length will also increase with the extent to which the costly information can be inferred from freely observable data. This is because better inferences will reduce the value of actually observing the information, resulting in less frequent observation and, hence, longer contracts.

What are the implications of these results for macro policy? It is clear that policies which somehow convey information about costly to observe parameters are the only ones with any chance of improving welfare given the above model. This observation has several interesting implications. First, to the extent that noisy government policies make it more costly to observe relevant parameters, such policies make people worse off (and increase contract length). Second, policies which aggregate information will tend to improve welfare. For example suppose that the value of  $\theta_t$  for one individual can be partially inferred from the "default" record of other individuals or from the observed value of  $\theta_t$  of other individuals. Further, suppose statistics are more costly for individuals to observe than their own  $\theta_t$ , but for the government the marginal cost of observing them is zero (for example, because they are needed for some other purpose anyway). If publication costs are small, it would be beneficial for the government to publish such data. Even if collection and publication costs exceed the benefits involved, some of the information can be revealed if some macro policy variables are correlated with this data. To be more specific, suppose that when default occurs, the individual involved takes some action such as reallocating his portfolio of securities. This will, in general, have effects on the money supply. By observing the money supply data and knowing the mecha-

nism by which the data are generated, one can infer something about the current economywide situation which, in turn, may be useful in estimating the current situation of one's trading partners.

The model suggests that the presence of useful information in aggregate policy data makes people better off than they otherwise would be and results in less rigid contract terms (but *longer* contracts). With regard to monetary policy specifically, our approach suggests that there are benefits of having the money supply be sensitive to the actions (for example, portfolio decisions) of the private sector. Our approach thus does not support the idea of 100 percent reserve requirements as has been suggested by Milton Friedman, since this would result in *M1* being independent of individual portfolio decisions. On the other hand, the model does support the use of a stable, nonstochastic rule (such as a fixed growth rate) for determining the stock of high-powered money, since noise in the determination of high-powered money would lessen the informational content of the various measures of the money supply. One must tread cautiously here, however, since other potential effects of having the money supply depend on individual decisions have been omitted from our analysis.

Our conclusion, that monetary policies are effective in contractual markets insofar as they alter the informational content of the money supply, is related to recent results by Laurence Weiss (1980). He shows that feedback policies are nonneutral even in a noncontractual world if individuals are differentially informed. The reason for nonneutrality in his model is that a change in monetary policy will alter the informational content of prices through changes in individual behavior. Since we allow contracts that are explicitly linked to the money supply (while Weiss does not have contracts), the informational connection is more direct, but in principle similar to that of Weiss' model.

In view of this connection, one might ask of what importance it is to include contracts at all in the discussion of monetary policy. There are two possible answers. First, the effects of monetary policy in a contractual

market may be distinctly different from its effects in a standard price-mediated market. Secondly, if the previous statement turns out to be incorrect, then the use of contractual models with their explicit price formation process may be an easier device for analyzing monetary policy.

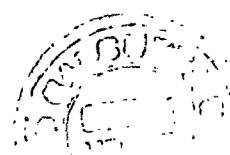
### III. Summary and Conclusions

In this brief report, we have examined the implication of some recent developments in contract theory for aggregate economic policy. We conclude that one role of aggregate policy is to convey information which would otherwise be more costly to obtain. This brings out the importance of how policy variables depend on individual-specific events and information as opposed to how they depend on past realizations of economy-wide shocks (as in Fischer, Phelps and Taylor, etc.). Finally, our approach suggests that the more information conveyed by policy variables, the less beneficial is privately obtained information. This will result in longer contracts, but with less rigid terms.

It cannot be overemphasized that our conclusions follow from a specific, highly simplified model. To analyze the informational effects of aggregate policies in any further detail, one must consider a model with many agents and some informational externalities. It would also be necessary to include in such a model features which allow these policies to have real effects apart from the informational effects (for example, as in Lucas). Although obtaining precise policy implications will be quite difficult when the effectiveness of monetary policy depends in part on its informational content, we feel that this is an important aspect and should be explored more fully.

### REFERENCES

- Barro, Robert J. and Grossman, Herschel, *Money, Employment, and Inflation*, Cambridge: Cambridge University Press, 1976.
- Dye, Ronald, "Contract Span and Specification Costs," Working Paper, Graduate School of Business, University of Chicago, 1981.
- Fischer, Stanley, "Long Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, 85, 191-205.
- Gray, Joanna, "On Indexation and Contract Length," *Journal of Political Economy*, February 1978, 86, 1-18.
- Harris, Milton and Holmstrom, Bengt, "On the Duration of Agreements," Northwestern University, Working Paper, Kellogg Graduate School of Management, August 1982.
- Keynes, J. M., *The General Theory of Employment, Interest, and Money*, New York: Harcourt, Brace and Company, 1964.
- Lucas, Robert E., Jr., "Expectations and Neutrality of Money," *Journal of Economic Theory*, April 1972, 4, 103-24.
- Muth, John, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, 29, 315-35.
- Phelps, Edmund and Taylor, John, "Stabilizing Powers of Monetary Policy under Rational Expectations," *Journal of Political Economy*, February 1977, 85, 163-90.
- Sargent, Thomas and Wallace, Neil, "'Rational' Expectations, the Optimal Money Instrument, and the Optimal Money Supply Rule," *Journal of Political Economy*, April 1975, 83, 241-54.
- Weiss, Laurence, "The Role for Active Monetary Policy in a Rational Expectations Model," *Journal of Political Economy*, April 1980, 88, 221-33.



# Is There a Monetary Business Cycle?

By CHRISTOPHER A. SIMS\*

There is a view that monetary policy is effectively summarized by the time path of a single variable, the money stock, and that erratic fluctuations in the money stock generated by erratic policy decisions are a major, even the principal source of business cycle fluctuations. This view, which I will call monetarism, is losing adherents among economists, for several reasons.

For one thing, rational expectations theory, having shown how the Phillips curve could emerge as a statistical regularity in an economy where it was not exploitable for policy purposes, can now do the same for Granger causal priority of money and money's strong explanatory power for future movements in real output. Theories of endogenous cyclical variation in money are gaining attention in part because of results which have begun emerging from the data as new statistical techniques are applied to new historical developments. Data from outside the United States or outside the 1950's and 1960's do not fit the predictions of monetarist theory.

## I. Causal Influence of Money as Statistical Artifact

In an earlier paper (1982), I showed how a stochastic equilibrium model could generate Granger causal priority of the money stock despite a kind of passive behavior of the monetary authority. Improving on that model slightly, it can be shown both how money might appear Granger causally prior with passive money, and how a policy of fixing the money stock might shift causal priority toward the interest rate.

First consider a simple model, containing identical infinitely lived agents maximizing

$$(1) \quad E \int_0^{\infty} [U(C_t) + H(M_t/P_t)] e^{-\delta t} dt,$$

where  $C_t$  is consumption at  $t$ , and  $M_t/P_t$  is

real balances at  $t$ , subject to

$$(2) \quad \dot{M}_t/P_t + \dot{K}_t = f(K_t) - C_t T_t + \varepsilon_t,$$

where  $T_t$  is taxes, and a given initial period per capita stock of capital  $K_0$ . The function  $f$  is the production function and the stochastic process  $\varepsilon$  represents random shocks to technology. I assume government policy imposes  $\dot{M}_t/P_t + T_t = 0$ . First-order conditions for solution of this problem with respect to  $K_t$  and  $C_t$  can be written as

$$(3) \quad U'_t = \phi_t,$$

$$(4) \quad -E_t \dot{\phi}_t = \phi_t (f_t - \delta),$$

where  $\phi_t$  is a random Lagrange multiplier and  $U'_t = U'(C_t)$  is the marginal utility of consumption at  $t$ . The left-hand side of (4),  $E_t \dot{\phi}_t$ , has a special interpretation here—it is the right derivative at  $t$  of expected future  $\phi$ . This may exist even when  $\phi_t$  is not a well-defined stochastic process, as when  $\phi_t$  is a Weiner process.

Equations (3) and (4) together imply  $E_t \dot{C}_t \neq \dot{C}_t$ , that is, that  $C_t$  does not have differentiable sample paths, so long as the future,  $\varepsilon_{t+s}$  for  $s > 0$ , is not known exactly at  $t$ . To see this, note that (4) with  $E_t \dot{\phi}_t = \dot{\phi}_t$  implies  $-\log \phi_t = \phi_0 + \int_0^t (f'_s - \delta) ds$ , and if  $U'$  is monotone in  $C_t$  there will then be a function  $g$  with  $g' > 0$  such that  $C_t = g(\int_0^t (f'_s - \delta) ds)$ .

Substituting this into the budget constraint (2) gives us

$$(5) \quad \dot{K}_t = f(K_t) - g\left(\int_0^t (f'_s - \delta) ds\right) + \varepsilon_t.$$

It is easy to check that this equation is "unstable."

If we have a path  $K_t^*$  which solves (5) for a given  $\varepsilon_t$  path, and we make a small, temporary increase in  $\varepsilon_t$ , then if  $K$  is unchanged at dates before  $t$ , the new path for  $K$  must differ from the old path by a positive amount which increases at least linearly in  $t$ , because

\*University of Minnesota.

the right-hand side of (5) is increasing in  $K_t$  and in  $K$ 's with earlier dates. Therefore a stationary  $\varepsilon$  process cannot generate a stationary  $K$  process in this system unless  $K$  has exact dependence on the sample path of  $\varepsilon$  in the future.

We can, however, have stationary solutions without dependence on the future if  $C_t$  does not have differentiable sample paths. For example, suppose  $U(C_t) = \log C_t$ , so  $\log \phi_t = -\log C_t$ . Then one solution to (2)–(4) satisfies

$$(6) \quad \log C_t = - \int_t^\infty E_t[f'_s + \gamma - \delta] ds + \Gamma,$$

which in turn is satisfied by the stochastic differential equation

$$(7) \quad d(\log C_t) = -(f'_s + \gamma - \delta) dt + (2\gamma)^{1/2} dW_t,$$

where  $W_t$  is a Weiner process.

A stochastic process like (7) does not have differentiable sample paths, and a  $C_t$  satisfying (6) cannot in general have differentiable sample paths if the future of  $f'$  is not known in advance. Processes of this type are locally unpredictable, in the sense that

$$(8) \quad \frac{E_t[(C_{t+\delta} - E_t C_{t+\delta})^2]}{E_t[(C_{t+\delta} - C_t)^2]} \rightarrow 1 \text{ as } \delta \rightarrow 0.$$

This means in particular that for data measured at small time intervals, the  $R^2$  in a regression of  $C_{t+1} - C_t$  on data available at  $t$  should be close to zero.  $C_t$  will therefore appear to be close to Granger-causally prior in systems of equations estimated with fine time unit data.

In any continuous time dynamic optimization problem with a productive investment technology, a control variable which can be moved without penalty on its derivative will tend to behave this way—attempting to track some function of expected future developments—and will therefore tend to have the property, which  $C_t$  has here, of being locally unpredictable. (See also my 1980c paper and Robert Hall, 1978.)

Hall showed in a discrete time model that consumption should be a martingale. His result depends on assuming a fixed real interest rate and on consumption varying only over a range within which the utility function is well approximated as quadratic. I have shown here that his result applies locally in time under more general assumptions.

The additive separability of utility makes the model dichotomize in a certain sense; we can solve (2)–(4) for  $C$  and  $K$  without solving for  $M$ . To solve for  $M$  we must use the first-order condition with respect to  $M_t$ , which is

$$(9) \quad H'_t = -E_t \dot{\phi}_t + \phi_t(\delta + E_t \dot{P}_t/P_t).$$

Using (3) and (4), this becomes

$$(10) \quad H'_t/U'_t = f'_t + E_t \dot{P}_t/P_t.$$

One possible monetary policy is to keep  $P_t$  constant, thereby keeping the nominal interest rate,  $f' + E_t \dot{P}_t/P_t$ , equal to the real rate  $f'$ . This makes (10) reduce to

$$(11) \quad H'(M_t/P_t)/U'(C_t) = f'(K_t).$$

Note  $K_t$  must be differentiable, since  $\dot{K}_t$  appears in the budget constraint (2), all of whose other terms have well-defined sample paths. Since  $C_t$  is not differentiable, and  $P_t$  is fixed, (11) implies that  $M_t$  must in the short run move directly with  $C_t$ , inheriting its non-differentiability and, therefore, its approximate Granger-causal priority with fine time unit data. The interest rate, depending directly on  $K_t$ , is differentiable and will not have this short-run volatility.

To institute this “price-pegging” policy, if the monetary authorities cannot simply intervene in “the” commodity market, requires that  $M$  be adjusted rapidly to the shifts in demand for it arising out of  $C$  shifts. The policy clearly puts a substantial burden on the monetary authorities.

Suppose instead the authorities peg  $M_t$  at a constant level. This leaves  $E_t \dot{P}_t/P_t$  in (10) and that equation becomes a stochastic differential equation in  $P$ , with  $C$  and  $K$ , determined by (2)–(4), as exogenously determined forcing functions.

Since  $H''$  is naturally taken to be negative, the partial derivative of the left-hand side of (10) with respect to  $P$  is positive, making (10) unstable in  $P$ . As noted above in discussing the solution of (5) for  $K$  in terms of  $\epsilon$ , this means that stationary solutions of the system without perfect foresight must make  $P$ 's sample paths non-differentiable. The nominal interest rate,  $f' + E_t \dot{P}/P = H'/U'$ , is likely to move directly with  $C_t$ , hence to be itself nondifferentiable.

In moving from a fixed  $P$  equilibrium to a fixed  $M$  equilibrium, we move from a situation with the rate of growth of  $M$  showing higher variance the smaller the time interval over which the rate is measured to one in which instead  $P$  shows this erratic behavior. The fixed- $P$  equilibrium requires the monetary authority to make rapid and accurate adjustments of the nominal stock of money to shifts in demand for money generated by real activity. The monetarist view might be that the monetary authority cannot do this. Note, though, that in looking at the historical record to decide if the authority has succeeded in doing this in the past, we should not expect that observing erratic and unpredictable movements in the money stock, highly correlated with subsequent movements of real output, is evidence that the monetary authority has failed. This model predicts just such statistical results precisely when the monetary authority is successful.

The fixed- $M$  equilibrium requires that the price level make rapid and accurate adjustments. My view is that the nature of contracting arrangements we observe in the economy suggests that a price level which is volatile and unpredictable over short time intervals would impose real costs which are ignored in this equilibrium model. These costs are probably big enough that, with a fixed- $M$  policy or something close to it, we see incomplete adjustment of prices, generating disequilibrium and inefficient responses in real activity.

To see why this is likely, consider for example what would happen in the model if information emerged which implied a quickly (but continuously) rising time path for consumption over some interval. This would mean  $U'_t$  was quickly dropping. With no

change in the nominal interest rate, the price level would have to drop as sharply as  $C_t$ . However the rapid fall in the price level would require a drop in the nominal interest rate, since the other component of the nominal interest rate,  $f'(K)$ , cannot change rapidly, and the drop in the nominal interest rate would further increase the demand for real balances, thereby further increasing the required price level increase. The result is that the price level must move down more sharply than consumption moves up to preserve equilibrium. In fact, if  $P$  and  $C$  are to have continuous paths, equilibrium will require that the rate of change of expected  $P$  and  $C$  both be differentiable functions of time, even though the levels of  $P$  and  $C$  are not differentiable. In other words,  $P$  and  $C$  levels both adjust so quickly to their new equilibrium levels that expectations of further change in the same direction are not justified. But suppose  $P$  is sluggish, moving in the right direction for equilibrium but not fast enough. Then an expectation of deflation will emerge, demand for money will increase faster than deflation is increasing real balances, and real activity is likely to be inefficiently depressed. In a pegged- $M$  world, good news about future consumption possibilities, not accommodated by the monetary authorities, paradoxically generates a deflationary contraction.

A modified nominal interest rate-pegging policy is possible. Fixing a constant nominal interest rate of  $\rho_0$  converts (10) to

$$(12) \quad \rho_0 = f'_t + E_t \dot{P}_t / P_t.$$

This equation does not yield a determinate value for the current price level. However, if the nominal rate is even slightly responsive to the price level, so the rate is set at, say  $\rho_0 + aP$ , with  $a$  positive, the price level is determinate. It is a function of the conditional distribution of future capital stocks and once again has nondifferentiable sample paths.

An interest rate pegging policy forces the price level to anticipate future levels of real returns to capital, while an  $M$ -pegging policy requires the price level to anticipate the gap between the value of the services of real

balances and the real return to capital. It is not clear a priori which is likely to produce a more erratic time path for the price level.

The model from which these conclusions are derived is special in some respects. Monetary policy could have real effects without the occurrence of disequilibrium if motivations for holding money are not well represented by separable utility, which they probably are not.

The time separability of the utility function in the model, though conventional, may not be realistic. The delayed response of output to surprise changes in the interest rate which appears in empirical work described later in this paper and elsewhere would be easier to explain if instantaneous utility depended on the first and second derivatives of consumption as well as on its level (see Alfonso Novales, 1982). If this were true, consumption would no longer be locally unpredictable. However, the denominator in (10), instead of moving directly with  $C$ , would instead move directly with the highest-order derivative of  $C$  entering the utility function, which would remain locally unpredictable. To change the model's conclusions about the short-run volatility of  $M$  under a price-pegging policy would require introducing some kind of inertia in holdings of money balances. But the costless adjustability of money balances is one of their principal distinguishing characteristics.

## II. Monetarist Alternatives

The model in the preceding section is one in which a strong statistical relation between fluctuations in money and future output fluctuations exists, yet stabilizing the rate of growth of money will not reduce fluctuations in output. The most explicit monetarist alternative is the recently influential rational expectations version of monetarism. This view insists that unpredictable shifts in the money stock are primarily generated by random policy decisions, not systematically related to contemporaneous private-sector developments. As represented in the work of Robert Lucas (1972) and others, models supporting this version of monetarism depend on introducing persistent informational asymmetries

across economic agents, a phenomenon not present in the models of this paper. A rational expectations monetarist perspective makes large unpredictable movements in the money stock evidence of bad monetary policy, and it implies that if the money stock could be made to grow in a smoothly predictable way, real fluctuations would be smaller.

Some monetarists accept the possibility that much or most variation in  $M$  is passive response of the money stock to business conditions, but argue that cyclical fluctuations in output would nonetheless be reduced if  $M$  were kept more stable. We do not have equilibrium models which explain why this should be so, but it is not an untenable view.

## III. Statistical Evidence

Table 1 shows standard deviations of annual changes in the *log* of the industrial production index and standard deviations of the change in the *log* of the money stock for five large, wealthy countries and two approximate decades (1960–70 and 1971–82). There is no relation across countries, and the two countries which reduced money volatility, France and the United States, showed the largest increases between decades in production volatility.

To retain the opinion that policy-induced reduction in the variance of annual rates of

TABLE 1—MONEY GROWTH VARIABILITY AND OUTPUT VARIABILITY ACROSS DECADES AND COUNTRIES

	1971–82		1960–71	
	$\sigma_M$	$\sigma_{IP}$	$\sigma_M$	$\sigma_{IP}$
United States	.0151	.0795	.0185	.0426
United Kingdom	.0557	.0642	—	—
Germany	.0422	.0682	.0247	.0638
France	.0281	.0564	.0516	.0328
Japan	.0707	.0447	.0462	.0440

Notes: All data are from OECD *Main Economic Indicators*. Listed statistics are standard errors of changes in logarithms taken over June-to-June 12-month intervals, except that for France, because of the drastic effect of political disturbances on June production during two summer months of the 1960–70 period, January-to-January changes were used for both  $M$  and  $IP$  for that period.

growth in money stock would reduce the variance of the rate of growth of output, one evidently has to interpret these data with some sophistication. It could be that there has been little policy-induced difference in  $M$  volatility across these countries or periods. Yet since there has been substantial variation of actual  $M$  volatility, the conclusion must be that much of  $M$  volatility is due to the influence of things other than policy. If these other influences are developments in the private sector, it suggests the importance of modeling why money responds to these influences before concluding that eliminating the responses would be a good idea. If these other influences are simply measurement error, then doubt is cast on the reasonableness of basing policy on targeting a variable so poorly measured.

Lucas's 1973 paper on the Phillips curve, it might be noted, made a point very similar to that being made here. He observed that variance of inflation rates bore no relation, in his international cross section, to variance in output; this suggested that policy-induced changes in inflation did not produce changes in output. I am making the same point with money stock in the role of prices.

Rational expectations monetarist models do conclude that it is not variation in the money stock itself which generates business cycle fluctuations, but unpredictable variations. One could imagine that the countries with large variances in the annual rate of growth of money stock nonetheless have more of that variance predictable than in the countries with smaller money volatility and large output volatility. This would lead to the conclusion, though, that targets for the rate of growth of money stock which vary a lot from year to year are as often associated with more predictable money stocks as not.

This paper's model has no difficulty rationalizing substantial changes in  $M$  volatility with the absence of associated changes in output volatility. Changes in monetary policy in this model leave the equilibrium stochastic process for output unchanged, yet could well have large effects on money volatility.

Even the within-country predictive relation of  $M1$  to output is inconsistent across countries. As I (1980a, 1982) and Robert

Litterman and Laurence Weiss, among others, have pointed out, the Granger-causal priority of money stock to output, which emerges strongly in money-output-price systems fit to postwar U.S. data, is much weaker in those data when interest rates are introduced into the system. Furthermore, results prepared for this paper (available from the author) are consistent with the conclusion of Michael Darby and James Lothian, using a different statistical approach, that in most countries, other than the United States, a relation of surprise movements in the money stock to real output is harder to discover.

#### IV. Conclusions

The matching of theory with evidence in this paper has been informal. Models of endogenous money of the type presented in the first part of this paper are difficult to estimate directly. Work now underway by Novales suggests that more formal testing of these models against data is possible, however, and he has constructed stochastic equilibrium models in which interest rates are dynamically related to output in the way that actual interest rates are in the systems estimated in this paper.

Despite the need for further work in this direction, the weight of evidence against the rational expectations monetarist view of the origin of the business cycle seems quite strong. There is little consistency across countries or time periods in the relation of money volatility or of money surprises to production. In all countries, money stock and output move smoothly together, money slightly in the lead, after interest rate surprises. That it is precisely these smooth, predictable joint movements in production and money which account for most of the correlation of those two variables seems to contradict the rational expectations monetarist position.

A monetarist view that accepted the endogeneity and predictability of money stock fluctuations, but argued that better results would be achieved if the authorities prevented these fluctuations, does not conflict as sharply with the evidence as does the rational expectations monetarist position. It is

hard to see, from this point of view, why prices if anything tend to increase following interest rate increases, while money decreases. If the authorities are paying too much attention to stabilizing nominal interest rates, one would expect a different result.

That the cross-country and cross-decade relations of money and production volatility are so weak seems incompatible with any version of monetarism which leads to a policy conclusion that we should focus policy on stabilizing annual growth rates of the money stock.

The conclusion then is that there is probably no monetarist business cycle. Of course, that is not the same thing as saying there is no influence of monetary policy on the business cycle.

#### REFERENCES

- Darby, Michael, and James Lothian, "Introduction, Summary and Conclusions from *The International Transmission of Inflation*," Working Paper No. 206, University of California-Los Angeles, 1982.
- Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis," *Journal of Political Economy*, December 1978, 6, 971-88.
- Litterman, Robert and Weiss, Laurence, "Money, Real Interest Rates and Output: A Reinterpretation of U.S. Postwar Data," Minneapolis Federal Reserve Bank Staff Report, 1983.
- Lucas, Jr., Robert E., "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, 4, 103-24.
- \_\_\_\_\_, "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, 63, 326-34.
- \_\_\_\_\_, "Understanding Business Cycles," in Karl Brunner and Alan Meltzer, eds, *Stabilization of the Domestic and International Economy*, Amsterdam: North-Holland, 1977.
- Novales, Alfonso, "A Stochastic Equilibrium Model of the Interest Rate," presented at the December Econometric Society meetings, 1982.
- Sims, C. A., (1980a) "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review Proceedings*, May 1980, 70, 250-57.
- \_\_\_\_\_, (1980b) "International Evidence on Monetary Factors in Macroeconomic Fluctuations," Discussion Paper, University of Minnesota, 1980.
- \_\_\_\_\_, (1980c) "Martingale-Like Behavior of Asset Prices," Discussion Paper, National Bureau of Economic Research, 1980.
- \_\_\_\_\_, "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 1:1982, 107-64.

## THE WORLD FOOD SITUATION

### Changing Trends in World Food Production and Trade

By G. EDWARD SCHUH\*

The key to understanding the world food situation lies in data on production of food, trends in prices of food, developments in the fertilizer industry, and trade. This paper contains a brief overview of each of these topics. For reasons of space I have chosen not to enter the debate on nutritional status. For a recent contribution to that debate, see Thomas Poleman.

#### I. Trends in Production

I have chosen to focus only on the latter half of the 1970's because so much of the recent debate on the food crisis was engendered in the early 1970's when there was a surge in international commodity prices.

The data tell an impressive story. By the end of the 1970's, food production for the less developed countries as a whole had increased 58 percent over the 1961-65 base period, compared to 42 percent for the developed countries as a whole. Africa was the poor performer among the continents of less developed countries, with food production having expanded only 32 percent over this period. However, food output in East Asia had expanded by 75 percent, by 73 percent in Latin America, and by 68 percent in West Asia. In Pakistan, food production doubled over this period, and it almost doubled in Brazil. (For more disaggregated figures, see D. Gale Johnson, 1982a.)

Even if one takes account of the more rapid growth in population, the performance of the less developed countries as a whole is still impressive. The annual growth of food production per capita in the low-income countries as a group was 0.7 percent during

the 1970's. The total masks a great deal, however. Per capita production of food declined by some 9 percent in Africa, and the increase for South Asia, which includes some 65 percent of the population of the low-income market economies, was only 8 percent. Bangladesh actually experienced a significant decline in per capita production, and per capita production also declined in each of the countries of the Caribbean in this same time period.

Although these data suggest a rather mixed and uneven performance on the part of the food economy, they are by no means consistent with any inference that we are running out of production potential, or that the less developed countries were unable to feed themselves. The increase in per capita production of 0.7 percent a year during the 1970's, although modest, was a solid achievement in light of the rapid population growth in these countries and in light of the discriminatory economic policies which most of these countries followed vis-à-vis their agriculture. Moreover, the performance of individual countries lends optimism regarding what might be done in the future. In the early 1970's, for example, India was an especial cause for concern. By the early 1980's, however, that country was actually exporting small surpluses of food grains.

The enigma, of course, is China, with its large population. Unfortunately, the data on this country are so precarious that it is difficult to know what to make of them. Johnson (1982a) has tried to disentangle these data. Among other things, he quotes a Chinese source which concedes that grain production per capita was about the same in 1977 as it was in 1955. But he also cites more recent official concessions which suggest that the performance of the Chinese food economy has been anything but good.

\*Professor and Head, department of agricultural and applied economics, University of Minnesota-St. Paul.

## II. Food Prices

The price of food is another measure of scarcity. If the demand for food was outpacing the supply, one would expect to see the relative price of food products increasing. In fact, it was the rapid rise in commodity prices in the early 1970's that gave rise to the world food scare.

Data on the real prices of wheat and corn in the United States provide a useful historical perspective. The price of grains is important since they are the primary food products of low-income people. Although far more rice is eaten as a main staple than wheat, the price of wheat is a reasonable proxy for the price of food grains. In contrast to rice, wheat is widely traded, and is often consumed in the countries that are major rice consumers. The price of wheat in the United States is also a reasonable proxy for the international price of wheat since the United States has long been an exporter of wheat and wheat products.

Since the 1860's, the real prices of corn and especially wheat have exhibited a long-term downward trend. Measuring in 1967 dollars, the price of wheat has declined from \$3.50 in the mid-1860's to \$1.70 at the end of the 1970's. While prices rose dramatically in the early 1970's, post-World War II price trend is also downwards, if anything at a more rapid rate than previously. It should also be noted that the peak in the 1970's did not ever reach the levels realized at the end of the war when wartime scarcity created the last significant markup in commodity prices.

The downward trending prices of staples provides some insight into the nature of the hunger or malnutrition problem. They suggest that this is an income or poverty problem and not one of production.

## III. Fertilizer

One of the concerns of the mid-1970's was that increases in fertilizer prices would dampen off the use of this modern input, thereby causing us to lose not only the contribution to production of that input, but also to fail to capitalize on the increased fertilizer-responsiveness of the new high-

yielding varieties. This concern about fertilizer was due to the dramatic rise in energy prices in the early 1970's; nitrogen fertilizer is based heavily on natural gas.

Fertilizer prices did increase in the first half of the 1970's—by 140 percent in terms of prices paid by U.S. farmers from 1970 to 1975. However, a longer-term perspective shows that the relative price of fertilizer rose only modestly. Table 1 presents indexes of prices paid by U.S. farmers for all production items and for individual components of the input mix. Compared to 1970, fertilizer prices at mid-1982 had increased less than 10 percent more than all production items. The bottom half of the table shows that this same pattern prevailed in the European Community.

The failure of fertilizer prices to rise as much as energy prices is explained in part by the fact that, important as natural gas is for the production of nitrogen fertilizer, the production process is still capital intensive. Hence the effect of energy prices alone was less than many observers expected. In addition, it appears that there was substantial growth in productivity in the fertilizer industry.

## IV. Trade in Agricultural Products

It was widely expected at the time of the world food crisis in the mid-1970's that the low-income developing countries would put inordinate demands on world food supplies due to their inability to feed their rapidly growing population. This is not what happened.<sup>1</sup>

Between 1969–71 and 1980, world grain imports increased by 118 million tons for more than a doubling of total imports. The centrally planned economies and the middle income market economies accounted for almost all of this increase, with the centrally planned economies accounting for the larger share (56 million vs. 37 million tons). These two groups together accounted for more than half the world's population (1,321 million and 933 million in mid-1978).

<sup>1</sup>For a more detailed analyses of international trade in cereals over this period, see Johnson (1982b).

TABLE 1—FARM INPUT PRICES, THE UNITED STATES AND THE EUROPEAN COMMUNITY  
(1970 = 100)

	1973	1975	1977	1979	1980	1981	Mid-1982
United States							
All Production Items	135	169	185	231	256	274	278
Fertilizer	117	250	208	225	279	300	304
Agricultural Chemicals	108	165	161	154	165	178	195
Fuels and Energy	110	169	192	263	362	410	384
Feed	159	185	185	203	228	248	235
European Community <sup>a</sup>							
Good and Services Used in Agriculture	130	169	211	234	257	291	—
Fertilizer	120	198	211	243	286	332	—
Energy and Lubricants	123	194	245	306	376	454	—
Feeding Stuffs	134	156	204	215	228	258	—

Sources: United States: U.S. Department of Agriculture, *Agricultural Prices*, various issues. European Community: Commission of the European Communities, *The Agricultural Situation in the Community*, annual reports for 1976, 1980 and 1981.

Note: This table is taken from Johnson (1982a).

<sup>a</sup>Data are for the nine members except that 1981 price indexes were based on differences between 1981 and 1980 prices for the ten members.

The low-income countries increased their imports only moderately over the decade. The same applies to the industrial economies, whose imports increased by only 22 million tons over the decade, or less than a fifth of the total increase in grain imports, while the industrial economies increased both their gross and net exports by a substantial amount, with the United States alone accounting for almost 70 percent of the total increase in grain exports for the decade.

The increased imports by the centrally planned economies reflected a number of factors. Bad weather has been the conventional explanation for the Soviet Union, although it is increasingly recognized that mismanagement of the agricultural sector has been an important factor. Holding the nominal prices of livestock products constant for the past two decades has compounded the problem.

In looking to the future, it is very likely that the Soviet Union will remain a major grain importer, unless there should be a major change in the economic regime. Imports by Eastern European countries have recently declined from their peak level of 1979–80 (from 17.5 million to 14 million tons). The

problem in this case has been the lack of foreign exchange.

China is the big enigma among the centrally planned economies, and in the global context as well. We understand only poorly what is going on in that country. Agricultural trade optimists see China as an enormous potential market for the future. The problem, however, will be foreign exchange. China's export capacity for other products would have to grow significantly if it were to have the foreign exchange needed to become a major grain importer. Moreover, other countries would have to be willing to accept those exports. Although somewhat equivocal, Johnson (1982b) judges that Chinese imports could remain at about the present level of 15 million tons a year. That is probably as good a guess as any.

Lesser developed, middle-income countries will probably continue to be strong importers in the decade ahead. How strong the demand from these countries will be, will depend on how strongly the international economy recovers and how broadly economic development spreads among the less developed countries. John Mellor expects the import demand from these countries to grow

very rapidly in the decade ahead. In my judgment whether this occurs will depend very much on government policies, and these are difficult to predict.

### V. Conclusions

To conclude, a few remarks on the role of the United States in the international food system are in order. The large export boom of the 1970's was in large part induced by a decline in the real value of the U.S. dollar in foreign exchange markets. In the beginning, this decline was a correction of the over-valuation of the dollar that had prevailed during the 1950's and 1960's (see my 1975 paper), but which became more severe in the late 1960's and early 1970's. However, during most of the 1970's, the dollar was weak in large part because as a nation we were subsidizing petroleum imports at the very time that OPEC was imposing large increases in petroleum prices.

President Reagan's deregulation of the petroleum industry, plus a renewal of the United States acting as central banker for the world, has changed all that. The U.S. dollar has risen approximately 25 percent in real terms over the last two years. That has choked off the U.S. export boom (see my 1982 study), with the value of exports now having declined over 10 percent. If OPEC should in fact break up and petroleum prices

continue to decline, we could see a continuation of that trend. Export supplies of grain will come in the future from a more diversified set of countries.

### REFERENCES

- Johnson, D. Gale, (1982a) "The World Food Situation: Development During the 1970's and Prospects for the 1980's," in Emery N. Castle and Kenzo Hemmi, eds., *U.S.-Japanese Agricultural Trade Relations*, Washington: Resources for the Future, Inc., 1982.
- \_\_\_\_\_, (1982b) "The Current World Food Situation," Conference Proceedings, *The Role of Markets in the World Food Economy*, Minneapolis, October 1982.
- Poleman, Thomas T., "World Hunger: Extent, Causes and Cures," Conference Proceedings, *The Role of Markets in the World Food Economy*, Minneapolis, October 1982.
- Schuh, G. Edward, "The Exchange Rate and U.S. Agriculture: Reply," *American Journal of Agricultural Economics*, November 1975, 57, 696-700.
- \_\_\_\_\_, "Agriculture in Transition," testimony presented to the Joint Economic Committee, U.S. Congress, April 1982.
- International Food Policy Research Institute, Report 1981, Washington: International Food Policy Research Institute, 1982.

# Food Prospects for the Developing Countries

By JOHN W. MELLOR\*

Current interest in food prospects for the developing world is based on a set of four relatively straightforward questions. Upon closer analysis, however, these questions prove to be successively more complex.

The first question is quite direct: Will food production in the Third World grow more rapidly than population? The answer seems to be a clear yes. Between 1961 and 1977, the growth of Third World production of major food crops averaged 2.6 percent a year, slightly higher than the 2.5 percent annual increase in population (Leonardo Paulino, forthcoming). There is every reason to believe that Third World food production in the future will continue to exceed population growth, since the processes for accelerated agricultural growth are now in place in so many developing countries and population growth rates are generally declining. The clear exceptions are Sub-Saharan Africa and the least developed countries (these are nearly synonymous). Even with a change in agricultural policies in these countries, there will be considerable time lags before food production growth rates exceed population growth rates.

The second question is more involved, and much more important with respect to its policy implications: Will ratios of food production self-sufficiency increase in the Third World? In the long run, self-sufficiency ratios in the Third World will indeed increase—but that is the long run of decades. In the short run of this decade and the next, these ratios will just as certainly decline, as rapidly accelerating growth in the demand for food in the Third World exceeds capacity to accelerate domestic production growth rates. This conclusion is reinforced by the tendency for accelerated food production growth to be

associated with forces that further accelerate growth in demand.

The third question is decidedly complex: Will the real price of food (defined in terms of relative shifts in the demand and supply schedules for food) shift upwards over the next two decades (as compared to the zero or slightly negative trend over the past few decades)? It is my judgment that it will. In the Third World, demand for food will clearly continue to shift more rapidly than supply. It is less certain that the forces in developing countries will overbalance the converse relationship for the developed countries.

The final question is the most far-reaching: What will be the impact of these forces on the nutritional status and the degree of poverty of low-income people? Since low-income people spend 60 to 80 percent of increments to income on food (see my 1978 article), food prices are a principal determinant of their real income and nutritional status. Increasing per capita food production and imports allow a rising number of people to eat better. Preliminary analysis of cross-section data for African countries show that as aggregate per capita food supplies rise, the proportion of malnourished children declines (see Shubh Kumar, 1981). Increased capital intensity and the dynamics of food production itself will raise real wages for much of the laboring class (see Uma Lele's and my 1981 article). But for some individuals and particularly for those in countries left out of development processes, the situation will be more difficult in the future than in the decades of the 1950's and 1960's. This, I should emphasize, is the note of pessimism in this paper.

## I. Past Trends in Food Crop Production, Consumption, and Trade in the Third World

The broad trends in food relations for developing countries are clear: food production growing more rapidly than population,

\*Director, International Food Policy Research Institute. I acknowledge the assistance of my colleagues at the International Food Policy Research Institute, particularly Richard H. Adams, Jr., Leonardo Paulino, J. S. Sarma, and Patrick Yeung.

but less rapidly than consumption; and exports growing, but less rapidly than imports. Between 1961–65 and 1973–78, net imports of major food staples by developing countries (excluding the People's Republic of China) increased from about 5 million tons to 23 million tons, an average rate of increase of 13 percent per year (Paulino).

Categorizing the growth in food imports by income growth rates illuminates trends. Countries with the lowest growth rates in per capita incomes (largely located in Africa) had the most rapid growth rate for net imports. This is the product of a very poor record in food production growth (well below population growth rates) and rapid growth of urbanization (itself a cause of the slow growth in agricultural production) based on large net capital inflows financed from foreign assistance.

The next three categories of per capita income growth have successively more rapid rates of growth of net imports. Presumably the 1–3 percent per capita income growth rate countries have relatively slow growth in demand for food and a foreign exchange allocation policy that restricts food imports. Those with successively higher income growth rates have a more rapid growth in demand for food and a trade regime that facilitates allocations to food imports.

Two elements of the recent past deserve special note: the very rapid growth in food imports by the fast income growth countries; and the association of the most rapid growth rates in food production with increased food imports.

During the period 1970–77, twenty Third World countries, containing some 700 million people, averaged growth rates of 4 percent or better in per capita income. Eight of these countries, all major oil exporters (Algeria, Indonesia, Iran, Iraq, Mexico, Nigeria, Saudi Arabia, and Venezuela), had an average per capita income growth rate of 5.6 percent per year. Such growth rates fueled an extraordinary rise in the demand for food in these countries, a demand that their still fledgling agricultural sectors were quite unable to meet. Thus, in the period 1970–77, food imports to these countries grew at the rate of 19 percent per year in real terms. In

the future, returns from high investment rates of the past and reduced future investment rates will both serve to sustain high per capita food consumption growth rates.

Perhaps even more symptomatic of the future, however, are the growth patterns of the twelve rapid growth Third World countries that are not major oil exporters. Brazil, Hong Kong, the Democratic People's Republic of Korea, the Republic of Korea, Malaysia, the Philippines, Singapore, Syria, Taiwan, Thailand, Tunisia, and Turkey had an average growth rate of per capita income of 5.6 percent for the same period. In these twelve countries, the demand for food rose by well over 5 percent a year. These countries are now, of course, experiencing a substantial slowdown in their growth rates as a result of the current recession. But they have developed the institutional base for sustained economic growth and can be expected to resume a rapid pace of growth as soon as the current recession has run its course.

Given the dynamic nature of these high growth rate countries, it is possible that some of them will be able to achieve the 4 or 5 percent annual growth rates in food production necessary to meet demand growth. However, the historical record suggests that most of these countries will not be able to achieve such food production growth rates.

More striking in terms of implications to trade in food is the finding that the sixteen developing countries with the fastest growth rates in basic food staples production over the period 1961–76 collectively more than doubled their net food imports (in tons) during the study period (see Kenneth Bachman and Paulino, 1979). From 1961–76, the average growth rate for basic food staples in these countries was 3.9 percent, but the high growth rate in food imports meant that their self-sufficiency ratio actually declined two percentage points. These data demonstrate that, while it is possible for rapid growth, low-income countries to achieve impressive increases in food production, it is still quite difficult for such production to keep pace with the rate of growth in demand for food.

It is important to note one further historical note that is useful for judging the future: improvement in crop yields was the main

contributor to the growth in food production in the developing countries between 1961 and 1977. Output per hectare of major food crops increased by 1.9 percent annually, and accounted for more than 70 percent of production growth, whereas increases in the harvested area averaged less than 1.0 percent a year, contributing the other 30 percent (Paulino). It is likely that the relative importance of yield increases as a source of growth will continue to increase. Yield increases are the product of technological change that requires complex institutional development and large numbers of trained people. These are processes that take time. In Asia, for example, technological change in agriculture requires large-scale public investment in irrigation, as well as a favorable price climate and an adequate infrastructure support system.

It is notable that in Africa where per capita food production has declined 15 percent since 1969–71, the median government expenditure on agriculture for the period 1963–73 was only 7.6 percent of the total government budget (see Norman Nicholson et al., 1979). Public sector allocations to agriculture are proportionately even lower in Latin America, but these are economies with much larger nonagricultural sectors (see Victor Elias, 1981). By means of contrast, in the Punjab of India, where the new agricultural technology has made a significant contribution to increased food production, the state government allocated to agriculture 11.0 percent of its expenditure.

## II. Phases in Food Demand-Supply Growth

The way in which the processes of economic growth accelerate growth in demand for food and causes food imports to burgeon during the developmental process is illuminated by depicting stylized phases of economic growth.

At an early stage of economic growth, people are very poor, desperately wishing to consume more food, yet unable to do so because of low incomes. In this stage, poverty causes high death rates and hence only modest rates of population growth, while per capita income grows hardly at all. The result is a 3 percent or less growth rate in effective

demand for food. That rate that can be met by more human effort on a slightly expanded land base. In this stage, population growth roughly meets its own demand for food.

As development occurs, the population growth rate increases. But, even more importantly, income begins to grow rapidly, and the two together increase the growth rate of demand for food by some 30 percent over the earlier phase. Such a rate of growth in food demand exceeds all but the most rapid known rates of food production growth. In practice, high income growth reduces two of the previous stages' sources of growth—expansion onto poorer, less-productive land area and the use of labor to intensify production at lower and lower returns to that labor. Thus prolonged and continued technological innovation in agriculture is needed in this stage, both to balance loss in production sources and to meet rapid growth in demand. Even 2 to 3 percent growth rates in land productivity are considered high. It is for this reason that most countries in the high growth, medium-income stage find it necessary to rely upon increasingly rapid growth in food imports to meet much of demand growth. It is only countries with unusual potential to expand onto high productivity land areas that can avoid this phenomenon.

In later stages, of course, population growth rates decline and growth in income begins to have little effect on demand for food. Meeting demand growth then becomes more manageable, particularly since by then food production growth rates have become institutionalized at high levels. It is in this stage that food imports become unnecessary and agricultural surpluses begin to accrue.

Thus it is increasing per capita income that is the dynamic factor underlying the growth in food demand in the Third World. However, it is important to realize that as demand for food rises in response to income, the relative composition of that demand changes over time. Rising income causes demand to shift to the more preferred cereals, and to highly income-elastic livestock products. The latter, in particular, become increasingly important in the consumption patterns of consumers, as evidenced by the fact that between 1961–65 and 1973–77, annual

meat consumption in the Third World increased at a 3.4 percent rate, significantly faster than population growth.

The rising importance of livestock products in rapidly developing countries plays a major role in restraining the decline in the overall income elasticities for basic food staples. The income elasticity of demand for livestock products remains fairly stable to relatively high income levels. This results in a strong derived demand for basic food staples, even as income rises. This phenomenon is reinforced by the fact that among livestock products, demand for pigs and poultry, both of which are produced at the margin largely on concentrate feed, grows most rapidly.

It is instructive to note here a peculiarity of the derived demand for feed for livestock and its effect on the aggregate income elasticity of demand for basic food staples. At low incomes, livestock commodities comprise a small budget share and hence the derived demand for basic food staples is quite small. As incomes rise, the income elasticity of demand for basic food staples for direct human consumption declines; but at the same time the income elasticity of demand for food staples for livestock consumption begins to increase. Initially, the base level of the derived demand is very small relative to direct demand, but, with sharply different elasticities, the relative weights change quite rapidly. Thus the income elasticity of demand for total food staples forms an S shaped curve, with the weighted average elasticity first declining, then rising, and then eventually declining again. It is in the period when the weighted average elasticity of both direct and derived demand peaks that developing countries move onto the international market for substantial aggregate imports of food. Given that imports are initially small relative to total consumption and hence highly leveraged, it becomes clear why analysts are normally caught unawares by the explosive growth in imports of basic food staples by developing countries after a period of slow growth in or even declining imports.

### III. Food Projections to the Future

On the basis of a straight-line projection by country of 1966-77 production and in-

come trend data, and UN medium population projections, a total net deficit of 75 million metric tons of major food crops in the Third World is projected by the year 2000 (Paulino). Although this represents only 5 percent of the total projected food demand in the developing world in that year, it is nearly three times the total estimated food deficit of these countries in 1977. As might be expected, the largest net production shortfall (65 million metric tons) is projected in those countries experiencing the fastest rate of income growth.

What caveats should we have in mind in using simple projections of the past to depict the future? First, the base period was one of unusually rapid income growth in developing countries. Certainly extrapolation of that period assumes not only an end to the current recession but additions to the ranks of fast growth countries to balance inevitable dropouts. These are hardly the days when such optimistic projections can seem realistic. However, large countries such as India and China have developed many of the characteristics thought to be precursors of accelerated economic growth. Further, few of the fast growth countries of the past decade have reached the stage of a major aggregate impact of growth in livestock consumption on demand for basic food staples.

Second, while it is unlikely that African food production growth rates will accelerate sharply in the near future, one may also question whether foreign resource transfers will permit continued growth at the depicted rates. On the other hand, the political pressures to preserve stability through foreign aid seem unlikely to change.

Third, the projected imports into the fast growth countries are so immense as to question the linear extrapolation. However per capita projected livestock consumption does not exceed levels of high-income Western countries, while cereals for direct human consumption are projected to decline.

Fourth, isn't it important to ask whether the prospect of such large food imports might not induce Third World policymakers to take steps to accelerate present food production growth rates? However, several of the fast growth countries have already achieved very respectable production growth rates, while

others have such rudimentary institutional structures as to delay the achievement of such growth rates for a long time.

It is considerably more difficult to estimate the ability of the developed world to respond to projected food deficits in the Third World. A simple projection of 1966-77 production and consumption trends projects a net surplus in the developed world of only 46 million metric tons of basic food staples by the end of the century (Paulino). However, this is the product of a huge surplus in the United States and a huge deficit in the Soviet Union. In contrast, projection of the longer period 1961-79 provides a net surplus of 196 million tons! Where one falls between these two projections is a function of such factors as: EC policy on prices; the extent to which U.S. technology generation can maintain high rates of yield increase; and, Soviet Bloc policies on meat consumption and rationalization of production policy. Perhaps it is not unreasonable to think that a downward trend in prices would bring little of the shift needed to generate larger net exports and an upward trend would bring a lot.

#### IV. Conclusion

Given the uncertainty, prudent action for developing countries is to search assiduously for cost-reducing technological change in agriculture. To the extent that policies in developed countries generate more than adequate surpluses, marginal investment in agriculture in developing countries will be justified by cost-decreasing technological change in agriculture. Because of the relative size of agriculture in developing countries, equity and income distribution concerns may still press for added emphasis on agriculture. Failure by developing countries to develop

policies and investment allocations favorable to expansion of agriculture runs the risk of suboptimal resource allocation if import bills and food costs are rising and the certainty of a narrow base of participation in development.

#### REFERENCES

- Bachman, Kenneth L., and Paulino, Leonardo A., *Rapid Food Production Growth in Selected Developing Countries*, Research Report No. 11, Washington: International Food Policy Research Institute, October 1979.
- Elias, Victor J., *Government Expenditures on Agriculture in Latin America*, Research Report No. 23, Washington: International Food Policy Research Institute, May 1981.
- Kumar, Shubh, "Nutrition Concerns in Food Policy for Sub-Saharan Africa," *Food Policy Issues and Concerns in Sub-Saharan Africa*, Washington: International Food Policy Research Institute, February 1981.
- Lele, Uma and Mellor, John W., "Technological Change, Distributive Bias and Labor Transfer in a Two Sector Economy," *Oxford Economic Papers*, November 1981, 33.
- Mellor, John W., "Food Price Policy and Income Distribution in Low-Income Countries," *Economic Development and Cultural Change*, October 1978, 27, 1-26.
- Nicholson, Norman K., Esseks, John D. and Khan, Ali A., "The Politics of Food Scarcities in Developing Countries," in Hopkins et al., eds., *Food Politics and Agricultural Development: Case Studies in the Public Policy of Rural Modernization*, Boulder: Westview Press, 1979.
- Paulino, Leonardo, A., *Food in the Third World: Past Trends and Projections to 2000*. Washington: International Food Policy Research Institute, forthcoming.

# Food Prospects in the Developing Countries: A Qualified Optimistic View

By MALCOLM D. BALE AND RONALD C. DUNCAN\*

The available evidence indicates that, in aggregate, the growth in world food production over the past two decades has more than kept pace with the growth in population. Improvements in per capita consumption among the developing countries have been widespread. The important exceptions have included many of the countries of Sub-Saharan Africa. Associated with this improvement has been the adoption and success of improved technologies, increased investment in infrastructure vital to increased agricultural production.

We believe it is likely that these improvements in food availability in the developing countries will continue, but that any such improvements would be considerably enhanced by widespread adoption in developing countries of pricing policies which remove the existing distortions under which agricultural production labors. Further it would enhance income distribution.

In this paper we illustrate the changes that have occurred in food consumption in developing countries over the past twenty years and present forecasts of food consumption growth which have recently been assembled in the World Bank. We are not directly concerned with the question of hunger, whether chronic or periodic. We agree with the view that hunger is not directly related to the level of world food availability, but is more a question of income level, or as Amartya Sen puts it, the "entitlement" to sufficient resources to purchase enough food to live. The many studies which the World Bank for one has done on the cost effectiveness of programs to meet chronic hunger among specific

groups and periodic hunger due to sharp reductions in food supplies have recently been summarized by Shlomo Reutlinger (1981/1982).

## I. Historical Growth of Income and Food Consumption

The last twenty years have been a period of substantial growth for some developing countries. How has this growth (or lack of growth) of the per capita incomes of developing countries affected food consumption in those countries? To obtain a broad picture of the impact, we have plotted elsewhere (1983) the growth in per capita calorie consumption of all food-stuffs against growth in income (*PPP*)<sup>1</sup> for the developing countries, for two periods, 1960-70 and 1970-79. In using this measure of improvement in food consumption, we recognize the difficulties associated with the measurement of "adequate" diets.<sup>2</sup> All that is being implied in our use is that at these levels of food consumption, growth in calorie consumption does represent an improvement in living standards.

We find that there is a reasonably strong positive correlation between per capita income and food consumption growth, with much less dispersion in the 1960's than in the 1970's. In the 1970's, a large number of developing countries experienced satisfactory per capita income growth (around 3 percent per annum), but little and often negative per

\*Country Policy Department, and Economic Analysis and Projections Department, World Bank, respectively. The opinions and interpretations presented here are our own, and may or may not correspond to those of the World Bank.

<sup>1</sup>Purchasing power parity (*PPP*) estimates of national income (i.e., estimated in terms of a set of international prices) are preferred to traditional exchange rate adjusted *GDP* estimates because they better reflect the purchasing power of income in each country. The *PPP* estimates flow from the United Nations/World Bank project on International Comparisons of Real Product carried out by Irving Kravis et al.

<sup>2</sup>For a critique of the estimation of basic food needs, see Nick Eberstadt.

capita food consumption growth (on average less than 1 percent per annum). Increased agricultural production instability in the 1970's is a likely cause for part of this behavior. (T. N. Barr shows that the variability of world food production was higher in the 1960's than in the 1950's, and higher again in the 1970's.) Reutlinger (1978) has shown that, in the face of reductions in domestic production, developing countries have been reluctant to compensate for the shortfall by increasing food imports. Moreover, artificially low food prices in developing countries, often aimed at the politically more powerful urban consumers, do not necessarily mean increased consumption. Besides the restrictions on imports, low producer prices mean lower incomes and hence lower consumption in the rural areas, where most of the population and most of the poor often reside.

What are the changes we are likely to see in the next twenty years in terms of the levels of food consumption reached in the developing countries? Further, what will the expected developments in food consumption mean for the future pattern of production and prices of foodstuffs?

## II. World Bank Forecasts of Food Consumption by 1995

The World Bank has recently carried out an exercise which in part led to forecasts to the year 1995 of world (and regional) production, consumption, trade and prices of many primary commodities of importance to developing countries. For the most part, these long-term projections were derived using comparative static equilibrium models, disaggregated by region, where prices are used to achieve a unique equilibrium solution. They are described further in our working paper.

The basic position that we take on projections of food consumption is that the amount and composition of food is, in the aggregate and in the long run, determined by aggregate demand. In this we differ from those who take a physical capacity-cum-productivity possibility approach. That is, we believe that the resources allocated to food production,

including resources allocated to productivity-enhancing research, are endogenous. We acknowledge that in some countries the food-producing sector is so large a proportion of total national product that it cannot be regarded as not being simultaneously determined with aggregate demand. Further, there are also distortions of prices which disturb food production from the levels that would otherwise be determined by aggregate demand. However, while these influences are important, often extremely important for individual countries, for this exercise we chose to assume that they are captured in the projections of income growth, since it is not possible to estimate their effects directly.

In looking at the level and composition of food consumption, our focus on aggregate demand implies that, both at the world level and at the country level, access to food is not determined by physical constraints on food production either within individual countries (because a country can import whatever food it can pay for) or at a world level (because production will respond to price incentives). Thus, it follows that we do not see the solution of any "food problem" at the world level or within a particular country as a question of overcoming food production problems (in a technical or agronomic sense) on a world basis or within a country, but as a problem of obtaining the maximum economic growth—within the economic constraints.

The long-term forecasts shown in Table 1 are conditional in nature. On the demand side, they rest critically on the assumptions made about the world economy in the 1980's and 1990's. The forecasts, moreover, are positive rather than normative. They are based on the most likely assumptions concerning government policies affecting production and trade, the likely market structures and demand conditions. Given the use of resulting price forecasts in project and balance of payments analyses by the World Bank Group, trying to determine what is most likely to happen, as opposed to what would happen if desirable changes in policies and market structure were to take place, becomes inescapable and appropriate.

TABLE 1—FOODSTUFFS—PROJECTIONS OF APPARENT CONSUMPTION, BY ECONOMIC REGIONS<sup>a</sup>  
(Million tons)

	Industrial Countries		Centrally Planned Economies		Developing Countries	
	1980	1995	1980	1995	1980	1995
Wheat	89.3(1.7) <sup>b</sup>	114.5(1.7) <sup>c</sup>	145.9(3.1) <sup>b</sup>	193.4(1.9) <sup>c</sup>	208.5(4.7) <sup>b</sup>	390.1(4.3) <sup>c</sup>
Rice	9.5(-0.6)	10.5(0.7)	14.5(1.9)	19.0(1.8)	236.1(3.0)	363.5(2.9)
Coarse Grains <sup>d</sup>	252.6(2.1)	270.4(0.5)	161.7(4.0)	249.0(2.9)	297.5(2.8)	481.8(3.3)
Sugar	25.1(1.3)	26.8(0.4)	17.7(2.3)	20.4(1.0)	54.8(3.6)	84.6(2.9)
Beef and Veal	20.6(2.1)	27.0(1.8)	9.5(4.1)	14.0(2.6)	16.5(2.6)	24.7(2.7)
Fresh Citrus Fruits	26.1(4.5)	34.0(1.8)	1.7(6.9)	2.7(3.1)	28.6(5.1)	48.5(3.6)
Vegetable Fats and Oils						
Soybeans (oil equivalent)	6.8(5.7)	9.6(2.3)	0.7(7.9)	1.2(3.7)	7.3(6.0)	16.2(5.5)
Palm oil	1.0(6.9)	1.3(1.7)	0.1(19.8)	0.2(2.9)	3.8(8.1)	8.5(5.6)
Coconuts (oil equivalent)	1.0(0.6)	1.3(1.7)	0.1(3.9)	0.2(2.7)	1.7(1.7)	2.6(2.9)

Source: FAO, *Production Yearbook 1981*, *Trade Yearbook 1981* (actual 1980 data); World Bank (projections).

<sup>a</sup>World Bank Classification of Countries, see *World Development Report 1982*. Note that China is included.

<sup>b</sup>The numbers in parentheses in this column are the actual growth rates (least squares trend) for the period 1961–80.

<sup>c</sup>The numbers in parentheses in this column are the projected growth rates (end-points) for the period 1980–95.

<sup>d</sup>Coarse grains here include maize, barley, oats, rye, grain sorghum and millet.

Observing the growth rates in world consumption for the period 1961–80 (Table 1), we see the declining importance of industrial countries in world consumption and the increasing importance of the centrally planned economies and the developing countries. The developing countries are expected to maintain their historical growth in grains consumption. This, together with the expected slower growth in population in developing countries should mean a slightly faster growth in per capita grains consumption than in the past twenty years. Bringing together the consumption growth rates of Table 1 and the population growth rates, per capita consumption of wheat, rice, and coarse grains by the developing countries was 2.5, 0.8, and 0.6 percent per annum, respectively, over the 1961–80 period. The projected per capita consumption growth rates for these grains in the period 1980–95 are 2.3, 0.9, and 1.3 percent, respectively. The much higher rate of growth of coarse grains is a reflection of the expected increase in the consumption of animal products.

In its major study of future food availability scenarios for developing countries, the FAO made the following projections. If, in the period 1980 to 2000, the *GDP* of develop-

ing countries grows at much the same rate as in the past twenty years, per capita food demand is projected to average 0.44 percent per annum growth. Given higher *GDP* growth, greater investment in agriculture, and freer trade, FAO projected that per capita food demand could grow by 0.75 percent per annum. The projections of the first FAO scenario appear compatible with our expectations of a slightly faster growth in per capita grains consumption in the next fifteen to twenty years.

### III. Price Developments

The cereals price forecasts from this exercise show a continuation of the long-term declining trend in real terms (figures are presented in our earlier paper). Even though incomes are increasing, and the growth of incomes in the developing countries is having a larger impact in terms of total food demand, food demand remains essentially price and income inelastic. Supply will respond to an increase in demand with improvements in technology. The result will be a fall in prices. The events of the past decade seem to bear this out. High prices early in the decade stimulated output which has resulted in low

prices since then. This is especially true for rice, the staple food of the major proportion of low-income people.

#### IV. Production and Yields

Grains occupy by far the largest part of land under agricultural production. Continuing increases in area harvested similar to those experienced in the past are unlikely.<sup>3</sup> Reliance on yield increases will probably rise. The potential for such increments exists and past experience, particularly in developing countries, is encouraging. If yields of wheat and coarse grains were to continue to grow at recent rates, projected production increases would be achieved with little increase in area harvested (with the exception of rice, where yields will need to and are expected to increase).

Yields are increasing at a constant or diminishing rate in industrial countries while they are increasing at an increasing rate in developing countries. In other words, in both rice and wheat, developing countries appear to be on the way to catching up to the industrial countries in yields. The exception to this encouraging performance by the developing countries has been in Africa. This is a very favorable indication of the future productive potential of developing countries; especially when the differences between yields in industrial countries and developing countries are considered. Even if developing countries never achieve yields of the same magnitude as industrial countries, merely approaching current yields in industrial countries would represent a substantial improvement and would have a significant effect on production. Since achieving yields similar to those now common in industrial countries involves adopting well-developed technologies, the technical challenge is not difficult. The main difficulty is an organizational one at the government level that involves providing correct incentives and removing obstacles to production increases.

<sup>3</sup>However, the FAO estimated that there was still considerable potential for area expansion in developing countries.

TABLE 2—NOMINAL PROTECTION COEFFICIENTS  
CALCULATED FOR GRAINS PRODUCTION  
IN DEVELOPING COUNTRIES

Country	Grain	NPC Estimate <sup>a</sup>
Africa		
Egypt	Maize	0.67
Ivory Coast	Rice	0.97
Kenya	Maize	0.91
Senegal	Rice	0.70
Sudan	Sorghum	0.50
Tanzania	Maize	0.13
Tunisia	Wheat	0.99
Zambia	Maize	0.62
Asia		
India	Rice	0.65
Pakistan	Wheat	0.76
Philippines	Rice	0.73
Thailand	Rice	0.58
Europe		
Turkey	Wheat	0.94
Yugoslavia	Wheat	0.38
Latin America		
Argentina	Wheat	0.64
Brazil	Rice	0.57
Colombia	Rice	0.92
Mexico	Wheat	0.89
Uruguay	Wheat	1.25

Source: World Bank.

<sup>a</sup>A value greater than 1.0 indicates a subsidy on production and a value less than 1.0 indicates a tax. These estimates have been made at different points of time and now may well be out of date; moreover, the estimates do vary widely from year to year within a country.

#### V. Other Factors

It is likely that within many developing countries, growth in food consumption has been negatively affected by distorted prices and income distribution which is biased against the rural poor. In Table 2 we present nominal protection coefficients (NPC) for 19 developing countries, calculated in most cases for the most important grain grown in each country. Bearing in mind the qualifications attached to these estimates, it is obvious that food production faces severe implicit or explicit taxes in many developing countries. It is our opinion that this factor has been one of the most important disincentive to the adoption of improved agricultural production performance in Africa, the region

of most persistent concern about human nutrition.

The creation of a "world food system" that has occurred since World War II has greatly alleviated the possibility of widespread food shortages. Currently, virtually the entire population of the world has access to the world food markets. Vastly improved communications, lower transportation costs, the construction of storage facilities, and the development of infrastructure have all contributed to the creation of this food complex. Food merchants receive worldwide market reports on a daily and sometimes hourly basis such that arbitrage largely equalizes the price of food commodities across the world (net of transportation costs, government intervention activities, quality differences, and the like). Because of these developments it is now possible to eliminate food shortages caused by natural events.<sup>4</sup>

Finally, it is well known that projections of economic behavior are notoriously unreliable (or are notoriously misinterpreted). All we know about the future is what we have observed in the past. We know that the future will be similar to the past because in the past the future has been similar to the past. Given this dictum, we interpret the information we have assembled in our earlier paper of over thirty years of declining agricultural prices and over thirty years of increasing crop yields (now increasing at an increasing rate in developing countries) as prima facie evidence of the robustness of the world food system, and of the likely continuation of such trends. We feel that it is incumbent on those who view the global food situation in a pessimistic way to provide a strong case of why trends that have been in existence for at least thirty-five years will be suddenly reversed. While we are cautiously optimistic about the continued im-

provement of food consumption and food output throughout the world, there is no room for complacency. As Johnson has observed, "if circumstances are to improve it is because efforts are made to make the improvement occur and at least some of the hinderences that exist, such as trade restrictions, low farm prices due to government constraints, and inadequate provision of farm inputs, are ameliorated" (p. 8).

## REFERENCES

- Bale, Malcolm, D., and Duncan, Ronald C. "Prospects for Food Production and Consumption in Developing Countries," Staff Working Paper, World Bank, 1983, forthcoming.
- Barr, T. N., "The World Food Situation and Global Grain Prospects," *Science*, December 1981, 24, 1087-95.
- Eberstadt, Nick, "Hunger and Ideology," *Commentary*, July 1981, 40-49.
- Kravis, I. S., Heston, A., and Summers, R., *International Comparisons of Real Product and Purchasing Power*, Baltimore: Johns Hopkins University Press, 1978.
- Johnson, D. Gale, "The Current World Food Situation," Conference Proceedings, *The Role of Markets in the World Food Economy*, Minneapolis, October 1982.
- Reutlinger, Shlomo, "Food Insecurity: Magnitude and Remedies," *World Development*, August 1978, 6, 797-811.
- , "World Bank Research on the Hunger Dimension of the Food Problem," *Research News*, 3, Washington: World Bank, Winter 1981/1982.
- Sen, Amartya, "Ingredients of Famine Analyses: Availability and Entitlements," *Quarterly Journal of Economics*, August 1981, 96, 433-64.
- FAO, *Agriculture: Toward 2000*, July 1979, Rome.
- World Bank, *World Development Report 1982*, London: Oxford University Press, 1982, 48-49.

<sup>4</sup>D. Gale Johnson observed this development to one of the authors.

## SEGMENTED LABOR MARKETS

### Labor Market Segmentation: To What Paradigm Does It Belong?

By MICHAEL J. PIORE\*

The dictionary defines a paradigm as "a model or pattern. In grammar, an example of a conjugation or declension, showing a word in all its inflectional forms." A paradigm is thus a practical example which perfectly illustrates an abstract principle. Its hallmark, in this dictionary definition, is the complete correspondence between the abstract and the applied, between theory and praxis.

The term was adapted to the discussion of scientific theory by Thomas Kuhn, and its current vogue in economics is attributable directly to his book. As Kuhn used the term, paradigm retains its double meaning of theory and of praxis, but the relationship between the two becomes ambiguous. Scientists believe, Kuhn seems to argue, that their inquiry is based upon a theory of how the world operates and how the investigation of its operation ought to proceed. But "normal science" is in fact largely a set of practices in which members of a given scientific community *customarily* engage. Students seeking to follow their professors in careers in which the latter will judge them and ultimately determine their fate are not encouraged to undertake projects which depart radically from those which their professors conceive for them. When suggestions for such projects arise, either from students or competing members of the community, they are as often treated by sarcasm and ridicule as the subject of seasoned discourse and debate. In this process, what students acquire is less an abstract understanding of what they are doing than a set of habits, or instincts, about what

constitutes a legitimate mode of inquiry or a plausible explanation. The disjuncture between theory and practice which arises in this way creates the potential for a "scientific revolution."

To ask in this context about the paradigm to which notions of labor market segmentation belong is thus to ask a question about the relationship between that mode of research and the theory and practice of normal economics. I never considered myself a revolutionary; indeed quite the contrary. But as an exponent of labor market segmentation in the community of "normal" economics, I can assure you that labor market segmentation does not fit the paradigm. The sentiments and reactions which Kuhn tells us greet the abnormal and *aparadigmatic* in a discipline, that is, fury, disdain, resentment, sarcasm, and condescension have definitely greeted labor market segmentation. This is a matter of fact; an observation about praxis. The question is then what aspect of the conceptual structure of the discipline accounts for this practical result.

In most discussions, labor market segmentation is contrasted with human capital theory, but this does not account for the hostile reception it has received. It fails, first, because the treatment accorded market segmentation by the profession is not that different from the treatment accorded human capital when it first appeared. Gary Becker describes his personal experience on this score in the introduction to his book of essays, *The Economic Approach to Human Behavior*, and the disdain and antagonism which he attributes to his colleagues are at least as great as anything experienced by the proponents of segmentation. But, perhaps more fundamentally, the two views are not, as we shall see shortly, necessarily in conflict.

\*Professor of economics, Mitsui Professor for Problems of Contemporary Technology, Massachusetts Institute of Technology.

The antagonism of conventional economics to labor market segmentation has more to do with where the observation comes from and how those who have responsibility for it sought to present it than in the existence of segmentation as a fact of nature. It has to do, in other words, with the practice of economics rather than with the discipline's theoretical content in the strict sense of the term. Two aspects of that practice are central. First, the manner in which it was "uncovered" involves approaches to empirical investigation which are excluded from conventional practice. The notion of labor market stratification emerged through "participant observation." The ideas were originally put forward by a group of us who encountered the labor market through participation in the civil rights movement and as advocates for the community based groups which grew up around that movement and President Johnson's War on Poverty. The ideas were an attempt to make sense out of the labor market problems as the people in these communities experienced them (or at least described their experiences) and to describe the labor market as these people saw it. To the extent that self-conscious, structured research was involved in the initial formulation of these ideas, the research was based upon relatively open-ended, unstructured interviews with the economic actors themselves.

This approach contrasts sharply with the practice of econometric estimation of *deductive* neoclassical models, using data gathered from highly structured interviews, the results of which are reduced, before they are introduced into the analysis, into continuous, quantitative variables. There is also a contrast in terms of the origin of the problems to which the research attempts to respond: conventional economists seldom derive their views about policy from the specific context in which social problems arise and programs are implemented. To find a precedent in economics for this kind of research, one has to go back to the old institutional labor economics of the 1930's and 1940's, and the generation of scholar-practitioners whose theory was an effort to organize their experience as arbitrators, mediators, and wage-control administrators, or to the early labor market studies of people like Reynolds or

Meyers whose research techniques in many ways simulated through interviews the exposure which they got in labor relations through direct participation. At the time in which stratification theories were being developed, the economics profession was in strong reaction to the "eclectic" nature of this research methodology and the *ad hoc* theories which it generated. Thus, however consistent the ideas might have been with orthodox theory, they were suspect because they were uncovered by unorthodox research practices and but for those practices might never have come into existence.

The second respect in which the notions surrounding labor market segmentation clash with the conventional paradigm is in the sharp discontinuities which they introduce into the world which theory has to explain. Conventional theory is infused by what one of my colleagues in physics calls an "aesthetic" of continuity and homogeneity. Exactly where this aesthetic comes from, how it is justified—indeed whether it is justified at all—is a topic too vast and complex for a paper of this kind. What I think is indisputable is that it is central to economics as praxis: the basic tools of theoretical analysis are applicable only in a continuous, homogeneous world, and the theories that are displayed in the classroom and constitute the standards of rigor and elegance against which students learn to judge their own work and those of their colleagues are theories about a continuous, homogeneous world. Any characterization that is sharply discontinuous and involves heterogeneous behavior is, on its face, intractable and unappealing. Were such a characterization to arise out of an empirical methodology that was nonetheless consistent with prevailing practice—whether it ever could arise in this way is another question—it might be treated as a puzzle for a clever theorist to solve. But, given that the empirical origins of labor market segmentation are already suspect, the theoretical aesthetic of the conventional paradigm strengthened the tendency to reject it out of hand. Segmentation is much more consistent with the theoretical aesthetic of Marxism economics, and it is, therefore, no accident that many of its chief exponents are radical economists.

A scientific paradigm, of course, exists not simply as theoretical and empirical practice. It has specific content as well. Here, I think it is much more difficult to know even how to broach the question, and the answer is a good deal more problematic.

At a certain level, the answer is clearly yes. Conventional theory does recognize sharp discontinuities between the labor force attachments of various demographic groups. Virtually all economists, for example, would accept a distinction between prime age working males, on the one hand, and women and youth, on the other. So long as the latter have a weak commitment to the labor market and a strong, inherent tendency to high turnover, one would expect distinct labor market institutions to govern their behavior. Add to that a certain variability in the stability of labor demand across different industries and occupations—a variation which one might well characterize, in accord with the prevailing theoretical aesthetic, as continuous—and one has a tendency toward exactly the dual labor market which was the fulcrum around which notions of labor market segmentation built. Most neoclassical theories of segmentation proceed along these lines. One can also build segmentation theories out of institutional imperfections in the labor market, out of the tendency for workers or employers to organize to protect their interests in the face of economic flux and competition. Since it is competition that generally enforces continuity and homogeneity in conventional theory, any abridgement of it will introduce the kind of discontinuous structure which notions of segmentation entail in a completely conventional way.

A stronger version of the segmentation hypothesis asserts that *behavior* differs systematically across market strata. Conventional theory has a little more trouble with this view. The convention is to assume that all workers are rational and that their labor market behavior is instrumental. The second assumption is clearly abridged in these versions of segmentation: arguably, the first assumption is abridged as well. On the whole, however, these stronger versions have simply been ignored.

It has been more difficult to ignore the notion of the *internal labor market*, which is

basically an assertion that in large territories of the labor market, job allocation and pricing are governed by institutional rules and customs which are only tenuously linked to rational, instrumental behavior or to competitive market forces, if they are so linked at all. But, even this challenge is limited, since the conventional theory does recognize certain building blocks of an economy, like the family or the firm. These basic building blocks are treated as coherent units, as if they were individuals. The theory generally ignores the way in which that internal coherence is achieved, and this is admittedly a weakness. But it need be no more of a weakness than the fact that convention takes the internal psychology of the individual as given and beyond the scope of its analysis. How big a weakness it *actually* is depends on how much of the action about the problems we wish to explain takes place internally, or how tightly constrained that internal action is by the external environment. The convention within the neoclassical paradigm is that it (the internal rules of the firm; the internal psychology of the individual) is either very stable, or so tightly constrained by the market that reference to the latter will explain its variability.

It is at this point, however, that the whole attempt to encompass notions of labor market stratification within conventional theory begins to break down. It breaks down because it depends upon bold assertions which one could in principle derive from a theory and investigate empirically. There *are*, moreover, theories about such things, and, on the whole, those theories are not consistent with the conventional paradigm, either as scientific praxis or as substance. Take, for example, the derivation of the dual labor market from the difference in labor force attachment among various labor force groups. The conventional focus upon women and youth is no accident: women and youth are biological categories. And biologically rooted behavioral differences combine relatively easily with an economic theory of social processes. But the question is not whether women and youth are biologically different from prime age males; the relevant question is whether their labor market behavior is a result of those biological differences. Since

that behavior varies historically, is currently undergoing significant change, and is demonstratively linked to social institutions like marriage, the laws governing military service, school attendance and the like, it seems doubtful that it can be biologically explained. That doubt is strengthened by the fact that all groups with a marginal labor force attachment to industry are not biologically based: worker-peasants, temporary migrants, even aspiring actors and artists play labor force roles similar to those of women and youth. To explain marginality, apparently one needs a *social* theory. And most social theories do not combine easily with the conventional paradigm.

Much the same can be said of institutional "imperfections" as an explanation of labor market segmentation. The conventional paradigm has no theory of such imperfections. In their face, it switches from a positive to a normative mode. It can explain behavior in their absence. And, in their presence, it prescribes their elimination. But it has no coherent theoretical story about where the imperfections came from and how (or indeed whether) they could in fact be eliminated. Nor does it have a good theory of how the economy behaves when imperfections are present. The lack of such a theory, moreover, is fundamental. Such imperfections invariably involve cohesive institutions, and hence any argument about the imperfections that one wants to get rid of would imply something about other cohesive institutions like the family and the firm, which are taken as the building blocks of economics and which the theory does not—indeed, could not—get rid of.

All of these questions push us to the limits of conventional theory and well beyond. The only conventional theorist who has actually tried to deal with them is Becker. That, I would maintain, is why Becker's work was initially treated with much the same ridicule and scorn as labor market segmentation. It is also why his theories of nonmarket social processes such as marriage which push conventional behavioral assumptions to their limit have not been taken seriously by most of his colleagues. To take them seriously would require either that we accept the full

implications of the conventional assumptions, or that we develop an alternative explanation. Most find the first obnoxious. The second is not possible within the conventional paradigm. It is precisely this dilemma that has led some of us to push for other ways of understanding labor market segmentation.

If ultimately labor market segmentation cannot be encompassed by the conventional paradigm, to what paradigm does it belong? At the core of labor market segmentation are social groups and institutions. The processes governing allocation and pricing within internal labor markets are *social*, opposed either to competitive processes or to instrumental calculations. The marginal labor force commitment of the groups which creates the potential for a viable secondary sector of a dual labor market is social. The structures which distinguish professional and managerial workers from other members of the labor force and provide their distinctive education and training are also social. To understand these phenomena, one therefore needs a paradigm which recognizes and encompasses social, as opposed to individual, phenomena. Here I see two alternatives.

One of these alternatives is Marxism. Its attraction is that it is a way of looking at and understanding the world in which society is a natural entity, and the social group, as opposed to the individual, is the focal point of analysis. Marxism, of course, is not *just* a paradigm in which human beings are understood in social terms. It focuses upon particular kinds of social groups, namely classes; it understands their nature and origins in particular ways, that is, the process of production; and it envisages only certain relationships among them, that is, subordination, exploitation, and conflict. These limitations seem especially severe in Marxism as it was understood and taught in the 1930's, and, as it continues to be presented in many liberal classrooms in the United States today. But they have not proved to be so debilitating in Marxism as a live and on-going framework of social analysis, and if one is willing to treat classes and their relationship to the productive process as entities which evolve and are transformed in history, which, I be-

lieve, is how Marx meant them to be treated, it is a framework which will encompass the observations about labor market stratification and render them intellectually tractable. As noted earlier, Marxism as a living intellectual enterprise is also more hospitable to the empirical approach to the external world which gave rise to the observations about labor market stratification in the first place. The advantages of this approach are evident in the work of the radical economists.

For me, however, Marxism as a paradigm has two major drawbacks. First is the enormous political baggage, the endless quarrels among Marxists themselves and between Marxists and conventional theorists, which the vocabulary of the Marxist paradigm draws one into. Most of these quarrels seem largely irrelevant to the intellectual purposes which one is about: many seem irrelevant to contemporary politics in general. But worst, it seems that to know where one stood on these issues, to know even whether one cares about them in the first place, one would have to resolve the intellectual puzzles which one wants the paradigm and its theoretical constructs for in the first place. And in resolving that prior problem, the quarrels are a continuing, and ultimately I fear, fatal distraction.

The second drawback of Marxism is that it fails to provide a bridge between the individual and the social. It seems to presuppose social man in much the same way that liberalism presupposes the isolated individual. And for somebody raised in the latter intellectual tradition, in a society like the United States which understands itself in individualistic terms, the failure is major.

The alternative to Marxism is to understand the society in terms of the nature of cognitive processes. Here, the critique of liberalism is that it views information and

thought, more or less as it views society, as consisting of discrete individual elements, which produce through aggregation the continuous homogeneous phenomena of social life. It fails to recognize that those discrete elements are processed in terms of some model, framework or "structure" of thought, and the same discrete "bits" lead to very different outcomes, depending upon what that interpretative framework turns out to be. Society, social groups, and institutions are part of the process through which that framework is generated. The need for such a framework is not a "bound" upon rationality, as my colleague Oliver Williamson would have it, but a precondition for it. Human thought, indeed human existence, is impossible without it. And, because it is so central, one must focus upon where the interpretative framework comes from and how it evolves. This notion of cognitive processes provides the framework for much of modern anthropology and sociology as well as developmental psychology and a good deal of modern linguistics. It is the core of what might be termed the "structuralist" paradigm. This is the basic paradigm to which Kuhn's theory of scientific revolution belongs. And I believe it provides the most promising framework within which to build an understanding of labor market segmentation. The scientific communities upon which Kuhn focuses are really "internal labor markets," segments of the market for professionals.

## REFERENCES

- Becker, Gary S., *The Economic Approach to Human Behavior*, Chicago and London: University of Chicago Press, 1976.  
 Kuhn, Thomas, *The Structure of Scientific Revolution*, 1962; 2d. ed., Chicago, 1970.

# Segmented Labor Markets in *LDCs*

By DIPAK MAZUMDAR\*

The largeness of the subject forces me to confine my discussion to the urban labor markets in *LDCs*, and furthermore to manual workers. I shall outline the process by which the urban market for workers of low skill tends to develop a sector of high wage, often accompanied by job security and fringe benefits unavailable to the large number of workers outside this sector. Furthermore, since the number of jobs in the high wage sector is limited, many urban job seekers with skills or human capital endowments similar to those employed in the sector have only limited opportunities of getting into it.

## I

The high wage or "formal" sector is easy to identify today by the operation of labor laws and/or unionism. But, while in a particular market the institutional factors might determine the point (or size class of enterprises) above which the formal sector extends, this sector cannot be said to owe its existence to institutional influences. First, the institutions have come into operation in a significant way only since World War II. But historical studies of labor markets show that the wage levels of regular workers in large-scale modern industry had been established well before the era of trade unions or government legislation (see, for example, my 1973 article, and Y. Yosuba, 1976). Secondly, there are many *LDCs* in which the substantive effects of trade unions or government legislation are limited even today, but wage levels are relatively high in the formal sector.

At the level of generalization, I would like to suggest a conceptual framework which seems to me to distill the experience of a number of *LDCs* regarding the emergence, consolidation, and exacerbation of labor

market segmentation. This framework might be viewed as describing both a sequential process of labor market development in a particular economy and a cross-section picture of different *LDCs* having reached the stages described in varying degrees.

## A.

The starting point should be a basic feature of rural-urban migration in *LDCs*—the distinction between different types of migrants with different supply prices. In particular, we should distinguish between individual migrants who came to the urban areas for long or short periods without breaking their ties with the rural economy, and family migrants who gave up rural residence (and activities) in a much more permanent way. The supply price of the latter will be necessarily at a higher level, and for several reasons. First, the loss of income in the family farm due to the absence of an individual may be considerably less than the total farm income, because other family members are able to substitute for his labor on the farm. This is particularly true when the absence is during the slack periods in agriculture. Secondly, the earner-dependent ratio for a family is significantly lower in the urban sector of *LDCs*, because of the more limited role of women and children in market activity in towns compared to the rural areas. Third, the cost of living in town for a family is higher not only because of the higher cost of housing, but also because the person who has migrated from the rural community with his family incurs the cost of finding for himself protection against old age, unemployment, ill health, etc.—protection which will ordinarily be provided by the Social Security system in a developed country, and which the rural family would provide the individual migrant. The institution of the extended family might reduce the cost for the rural born individual who migrates to town with his wife and

\*Senior economist, Development Research Department, The World Bank. The views expressed herein are my own; the usual caveat applies.

children, but it is unlikely that it will be for the entire amount.

Given this difference in the supply prices of individual and family migrants, if demand for labor in the urban market were undifferentiated, little family migration would take place as long as the operation of the agricultural economy allowed for a plentiful supply of individual migrants. This was indeed the experience of large parts of the African urban or nonagricultural sector for a long time before the end of the colonial era. In some *LDCs*, however, the employers in emerging modern industry saw the value of the higher efficiency of stable labor committed to industrial work over the major part of their lifetime. Wage levels were set at a high enough level to attract stable family migrants. Thus we find the emergence of a modern sector with wage levels that were perceptibly higher than the earnings of labor in those activities in which individual unstable migrants dominated (see my 1973 article). These activities—like casual labor—did not provide the same incentive for wage increase because of a weaker link between stability and efficiency. The higher wage level of stable migrants in the mechanism described so far could be interpreted as the reward for superior labor (from the point of view of employers) with a higher supply price.

### B.

I have so far discussed wage policy geared to creating a stable labor force by the formal sector as a whole, as though all employers were acting in unison. But, in a large urban labor market, individual employers will have their own policies of wage setting. The step from a wage policy designed to provide a labor force stable in terms of its urban residence to one aimed at keeping a firm-specific labor force is a short and perhaps inevitable one. This creates its own momentum for further wage increases. A profit-maximizing firm has an incentive to increase the wage of its own labor force as long as the increase in wage leads to a more than proportionate increase in efficiency. With a firm-specific labor force, the likelihood of the "least cost" wage being established at a relatively high

level is increased for several reasons. First, the firm is dealing with a body of workers separated from the rest of the work force, so that the benefits of the wage increase are not shared out among a large number, as will happen if they worked for a number of employers over a period of time. Secondly, the employer-employee relationship takes on some of the characteristics of an implicit contract with the understanding that the employee would achieve a certain level of efficiency and that employers would not pass on short-term fluctuations in demand by cutting wages. Thirdly, management costs are smaller the smaller is the work force one deals with, and hence there is a built-in incentive for employers to increase wage rates rather than hire extra workers as long as efficiency responds to wage increase.

The mechanism discussed here establishes wages in the formal sector at levels even higher than the supply price of stable labor, but efficiency wage is not correspondingly higher. In the neoclassical model, this process could be viewed as increasing wages to select a body of high quality labor. The relatively high wage, in other words, merely represents the rent accruing to superior (and scarce) labor. The alternative view to be considered is that the causal mechanism does not run from efficiency to wages, but the other way round. Suppose stable labor is in elastic supply to the formal urban sector at a certain wage. It is not in scarce supply at this wage. A certain body of workers is selected (more or less at random) to provide a stabilized work force. The subsequent increase in wage is not due to the prior scarcity of workers of a certain quality. It is due to the pursuit of a high-wage policy within a firm dealing with an exclusive body of workers—which produces net profits to be shared between management and workers.

### C.

The third stage in this schematic presentation of an urban labor market in *LDCs* is the opportunity offered by a stabilized and firm-specific labor force to the coming of unions. The process described above suggests that stabilized labor in the formal sector

might develop into a group which does not compete with and is not replaceable by the general mass of urban labor. Such a body of workers can be and often is organized. But the point to be emphasized is that unionism is established as a natural consequence of the labor market segmentation produced by the previous stages, rather than being a cause of the segmentation. In fact, in many situations employers are themselves known to have encouraged "responsible" unions as a protection against troublemakers. A point of some general applicability is that unions are often built on support from the government, who are also known to pursue labor policies on fringe benefits, job security, etc., and serve to enhance the privileged nature of formal sector employment. Government industrial policies in LDCs often create opportunities for high value productivity in modern industry through protection, cheap credit and so on. Interests of political stability demand that part of these "rents" created by public policy for private industrialists are shared with their workers. Pressures for such sharing are stronger when the employers are multinationals.

## II

Studies of wage differentials between the formal and informal sectors have been done within a framework of human capital model of wage determination. First, we want to see if there is a "net" difference in wages after controlling for education, experience and other human capital endowments. Second, a point to be investigated is if education or experience are being used as screening devices for selection of workers in the high-wage sector. Third, a major point of interest is if returns to education and experience are themselves significantly higher in the formal sector of the labor market.

An example is my 1979 study of the Bombay Labor Market. I undertook a multiple classification analysis (a form of the analysis of variance) of the determinants of the earnings of a sample of 5,000 wage workers in Bombay City in which the sector of employment (casual, small enterprise, and factories distinguished by three employment size

group) was included along with human capital variables (education, age, knowledge of English, and training). All the variables were highly significant and the model explained no less than 68 percent of the variance. But the important point was that the "sector of employment" variable turned out to be the most important explanatory factor, measured either by the ranking of the various variables in the order of explanatory power, or the spread of earnings associated with each category of variables after controlling for the others. Thus workers in factories employing 500 or more workers earned two and one-half times the casuals, after controlling for other factors, while the earnings of workers with postsecondary schooling were only 40 percent higher than those of illiterates.

It is also important to note that the increase in "net" earnings became stronger with the size class within the factory sector. Workers in factories employing 500 or more workers earned 60 percent more than workers in factories of 10-99 workers. (The wage difference between the latter and workers in small establishments with less than 10 workers was about 40 percent.) The result underlines the limited importance of purely "legal" factors in creating wage differentials since the Factory Acts cover most workers in enterprises with more than 10 employees.

The model was tested for the two-way interaction of the explanatory variables, and it was only the interaction between the size of the firm and the age of the worker which was of significance. The educational distribution of the workers in the different sectors of the labor market was very similar. Returns to the various educational levels, although significant in each sector, did not differ very much as between sectors.

The age-size interaction is important in showing the predominance of young workers in the small scale and casual sectors of the market. This is probably the result of a significant rate of return migration to the rural areas of workers staying in these sectors. However, the hypothesis that age-earnings profiles are flat in the informal sector of the labor market was rejected. Among those who continued in the informal labor market,

particularly in small establishments, earnings responded to age as much as in the factories.

The role of the employment size of firm in accounting for wage differentials (after allowing for human capital factors) has been noted in many economies in Asia, Africa, and Latin America. The quantitative importance of this factor would obviously vary. In terms of the conceptual framework of Section I, it would be large in economies where a plentiful supply of rural-urban circulatory migrants keep up the competitive pressure on wages in the small-scale and casual sectors. At the other end of the scale, the importance of large-scale firms, making use of capital intensive modern technology, could also exacerbate the size related wage differential. In Japanese studies, where this phenomenon has been recognized for a long time, it has been related to the process of technological diffusion. "Dualistic" industries defined as those showing wage differentials by size of enterprise have been industries which had been showing rapid change with borrowed foreign technology—and the list of such industries was different in different periods of Japanese economic history (see Yosuba).

### III

Even if a high wage formal sector is identified in the urban economy, it does not follow that the labor market is segmented. As noted earlier, there is a significant difference in age distribution between the Bombay factory sector and the other sectors. It might be hypothesized that the low-wage sectors are staging posts in the process of entry into factory jobs, as in the Harris-Todaro class of models.

The significance of this point turns very much on the empirical evidence on the extent of "graduation" from the informal to the formal sector. The two labor market studies I have done—one in urban Malaysia and the other in Bombay—both show graduation does exist, but is limited. In Bombay City, the proportion of large-scale factory workers at the time of the survey whose first jobs were in other urban wage sectors was about 25 percent. In urban

Malaysia, the proportion was a little higher, partly because the survey included movement out of self-employment. These figures, of course, represent the average experience of a large number of cohorts over varying periods of time. In his study of migration into Delhi, Biswajit Banerjee compared the proportion of new migrants who had entered the informal sector in a particular year but had moved to the formal sector within a 12-month period with the proportion of new arrivals who had found jobs in the formal sector directly. His figures showed, for example, "that in 1967 new arrivals were at least four to six times more likely to get formal sector employment than those who had entered the informal sector in 1966." The Asian evidence points strongly to the conclusion that the market for recruitment to formal sector jobs is located much more in rural areas than in the urban informal sector, as implied by the graduation hypothesis. This is particularly surprising because the rate of growth in employment in the formal sector is generally much longer than in the informal—and a large wage gap exists in favor of the former.

The reasons for this are partly on the supply and partly on the demand side of the labor market. On the supply side, the sustained impact of return migration and low wages on the potential efficiency of a worker seeking entry into the formal sector is significant. On the demand side, the value attached by employers to the social cohesion of a firm specific labor force leads them to depend on existing employees or their plant level supervisors to introduce new applicants for vacancies. Studies in India and Africa have repeatedly noted the importance of kinship ties in the recruitment process so that we end up with what Poppola noted in his study of Ahmedabad factories, a "de facto closed shop system."

This process of recruitment has the strong implication that the traditional hierarchical systems of caste relationships in rural labor markets would be carried over into the urban labor market. John Harris in his 1982 study of the Coimbatore labor market in India found that the dominant agricultural castes of the region were most strongly represented

in the regular work force of large engineering firms, whereas general casual labor in the town was dominated by members of the scheduled castes (who also constitute the bulk of landless agricultural labor).

#### IV

I conclude with some reflections on the welfare implications of the type of labor market segmentation above. I view the problem of large wage differentials within the urban manual work force more as one of distribution of income than as creating "distortions" or inefficiency in the neoclassical sense. As I have argued, the difference in efficiency wages (and hence of labor costs) between different size classes of enterprises could be much less than the observed wage gap. Nor can it be maintained that the creation of a stable, firm-specific labor force with virtual security of tenure in the formal sector creates rigidities in the system. The principal method by which large firms take care of fluctuations in labor requirements (and also cut down on the costs of fringe benefits and costs of recruitment) is the use of a body of "temporary" or "casual" workers. They are found in most *LDC* labor markets today, and are well known to students of Japanese labor history.

The concern with the distribution of income has both a static and a dynamic aspect. At any point in time, labor market segmentation, as I have discussed it, creates a privileged (and to some extent closed) class of workers who share in the relatively high net value added created in the formal sector, but are a minority of the urban work force. The course of development of the urban economy, in the recent history of many *LDCs* have added an extra dimension to this problem. A rate of growth of employment in the formal sector in excess of the rate of growth of the labor force would provide the possibility of an increasing proportion of the urban labor force being included in the high productivity formal sector. Typically, however, *LDCs* have had a rate of growth of employment in the formal sector well below the growth rate of the working population with

employment lagging significantly behind the growth rate of value-added in the formal sector. In some situations, the increase in the "least cost" efficiency wage of labor with technological progress is central to this development. In other situations, the market structure in which formal sector producers operate with protection from foreign competition, encourage inflation of wages and prices over time rather than translation of higher productivity into falling prices. In still other cases, direct institutional effects on the urban labor market have reinforced the market-determined developments stressed so far. Examples are the successive increases of minimum wages applicable to the formal sector in several African economies—which were originally intended as instruments to stabilize the migrant urban labor force, but overshot the mark. In many *LDCs*, the public sector is a large employer of labor outside agriculture, and has yielded to pressures to increase wages of the existing work force which, of course, reduces its potential for hiring new workers with its budget constraints. The consequence of all these developments is that the urban economy is characterized by widening wage differentials between the formal and the residual sectors with a declining proportion of the urban labor force employed in the former. What happens to income distribution in its lower reaches depends crucially on the dynamism of the self-employed sector of petty producers and traders—which often shows high returns to small doses of capital and entrepreneurship, and which may provide an alternative to job seekers not able to get into the high-wage sector.

#### REFERENCES

- Bannerji, Biswajit, "Some Aspect of Rural-Urban Migration in India: A Case Study of Delhi," unpublished doctoral dissertation, Oxford, 1981.
- Harris, John, "Small-Scale Production and Labor Markets in Coimbatore," *Economic and Political Weekly*, June 1982, 993–1002.
- Mazumdar, Dipak, "Labor Supply in Early

Industrialization: the Case of the Bombay Textile Industry," *Economic History Review*, August 1973, 26, 477-96.

\_\_\_\_\_, "Paradigms in the Study of Urban Labor Markets in LDC's: A Reassessment in the Light of an Empirical Summary in

Bombay City," World Bank Staff Working Paper No. 366, December 1979.

Yosuba, Y. "The Evolution of Dualistic Wage Structure," in Hugh Patrick, ed., *Japanese Industrialization and its Social Consequences*, Berkeley, 1976, 249-98.

# The Internalization of Labor Markets: Causes and Consequences

By BERNARD ELBAUM\*

A major area of debate in labor economics concerns the causes and consequences of the pricing and allocation of labor, within large firms, by internal administrative rules. At the center of controversy is the issue of whether or not labor markets are segmented between primary firms (offering the characteristic job stability and promotion opportunities of "internal labor markets," along with high career wage earnings) and secondary firms (offering unstable, dead-end jobs and low career wage earnings to equally capable employees).

By contrast, there has been relatively little debate over the related issue of competitive segmentation between enterprises with internal labor markets. Yet the significance of this latter sort of segmentation is suggested by the frequent practice of firms throughout an industry of filling all job vacancies, other than certain "port of entry" positions, by internal promotion.

Many jobs apparently require skills that are entirely or partially industry-specific. Production job skills seem especially liable to be industry-specific, as illustrated by the examples of coal miner, steel roller, and airline pilot. When rigidly maintained throughout an industry, internal promotion practices insulate the wages of such jobs from competitive market constraints. Relative occupational wages may then be substantially affected by bargaining, managerial policies, and custom without causing shifts in occupational employment patterns, as firms, like prospective employees, mainly concern themselves with expected career wage offers, and the implied average establishment wage. As a result, occupational wage rates may be very different from marginal products. In many industries, lack of external hiring over long time periods in itself provides *prima*

*facie* support for the salience of this possibility, as it is a sign of inoperative market policing mechanisms.

Empirical assessment of this latter type of segmentation involves three main questions: 1) the causes of labor market internalization; 2) the determinants of enterprise wage structure; and 3) the reasons why firms rely *exclusively* upon internal promotion. Below I review the distinctive implications for these questions of the neoclassical, radical, and institutional approaches to internal labor markets. I then evaluate how well these different perspectives can account for the findings of a detailed case study of internal labor markets and wage structure in the U.S. and British iron and steel industry. The conclusions offer my own interpretation of why internal promotion practices often are invariably maintained, and the operative constraints on wage structure.

## I. Alternative Perspectives on Internal Labor Markets and Wage Structure

From the neoclassical viewpoint, internal labor markets promote economic efficiency, and are explained by enterprise-specific human capital, which may take such forms as on-the-job training, or the screening of workers of differing ability by direct observation of job performance. With workers employed in the same enterprise for their careers, occupational wage rates are indeterminate. However, wage indeterminacy is bounded by the size of enterprise-specific investments in human capital. Occupational wage rates can be no less than prevailing wages for workers with comparable general training, net of search and screening costs, and no more than the sum of prevailing wages and the additional value productivity of incumbent workers in the enterprise. The expected present value of career wage earn-

\*Assistant professor, department of economics, Boston University.

ings must also accord with competitive standards. If worker risk preferences are added to the model, a unique wage structure may be derived as optimal for the firm, though problems of bargaining indeterminacy remain.

Employing a somewhat different model, Oliver Williamson, Michael Wachter, and Jeffrey Harris (1975) have attempted to reach analogously neoclassical conclusions. According to these authors, internal labor markets provide a system of collective regulation which resolves otherwise difficult and indeterminate bargaining between the firm, and work groups with enterprise-specific skills. In this situation of small numbers exchange, the effectiveness of individualistic employment contracts is problematical. Managers, being only human, have a limited capability for acquiring, handling, and communicating requisite information about ever-changing conditions. Work groups, for their part, can opportunistically obstruct contractual agreement or enforcement by collectively withholding or distorting relevant information. By shifting to the administrative arrangements of internal labor markets, the argument goes, the firm attaches wage rates mainly to jobs rather than workers, and discourages small-group bargaining, which frequently is replaced with job evaluation. Furthermore, "internal promotion ladders encourage a positive worker attitude towards on-the-job training, and enable the firm to reward cooperative behavior." The resulting wage structure, it is concluded, "reflects objective long-term job values rather than current bargaining exigencies" (p. 276). However, just why internal labor markets should necessarily yield this outcome is unclear. The preceding argument only appears to support the comparatively uncontroversial conclusion that, in light of managerial limitations and work group bargaining leverage, internal labor markets may be less costly than alternative enterprise policies.

Similarly, from the radical perspective, internal labor markets are functional for the firm because they promote employee identification with enterprise goals, and effective managerial control over work standards. Katherine Stone (1974) contends that U.S.

iron and steel firms, with these ends in mind, unilaterally imposed highly differentiated job and pay ladders around the turn of the century, although revolutionary technological change had by then rendered occupational skill distinctions "virtually meaningless" (p. 73). By contrast, other writers attribute the origins of internal labor markets to union instigation and to compromise arrangements negotiated between union leadership and management (David Gordon, Richard Edwards, and Michael Reich, 1982). But whatever their origins, in the radical view, internal labor markets generally embody a "bureaucratic" form of management control over workers, with little or no basis in technology. Occupational wage inequality is, by implication, largely an artificial construct, constrained by managerial calculation of the costs and benefits of pay differentiation. However, radical literature has paid little explicit attention to why wage structure varies between firm job ladders, firms, and industries; or to how wages are sheltered from competitive constraints.

From the institutional viewpoint, unlike the previous two views, job and pay structures within internal labor markets are significantly affected by ongoing bargaining between the firm, and workers and their organizations. Although the factors cited by other viewpoints may also be relevant, at heart internal labor markets reflect a quasi-legal legitimacy attached to workers' desires for security and advancement, backed up by the ability of work groups to inflict damage upon the enterprise if customary norms are violated. Within the spirit of the institutional approach, the rigidity of promotion rules should afford latitude for deviation from competitive outcomes. Recent empirical studies confirm this expectation. In particular, Richard Freeman (1982) finds the predominant effect of trade unions on the dispersion of earned income comes through their impact on internal wage structure. Descriptions of the pay policies of large establishments in Japan, or of large nonunion companies in the United States, further suggests that their wage structures depart from any close correspondence to marginal productivity relations.

However, institutional writers have expressed no consistent viewpoint on the type of labor market segmentation presently at issue. In Peter Doeringer and Michael Piore's treatment (1971), as in neoclassical analysis, indeterminacy in wage structure is bounded by the extent of enterprise-specific human capital. On the other hand, Robert Raimon (1953) views wage indeterminacy as the result of the prevalence, throughout the economy, of minimal training requirements, and of entry barriers for semiskilled employment posed by internal promotion practices. In Raimon's eyes, market wage rates for unskilled and skilled craft occupations set the bounds of indeterminacy for the wage structure of semiskilled job grades. However, Raimon does not explain why barriers to labor market competition should be so effective, or, in particular, why firms throughout the economy should refrain from poaching semiskilled workers employed elsewhere at relatively low occupational wage rates.

## II. The Case of Iron and Steel

The different implications of these perspectives may be highlighted by rephrasing the questions for empirical analysis as follows: 1) What roles were respectively played in establishing internal labor markets by efficiency considerations, management initiative, and pressures from the workforce? 2) To what extent is wage structure based upon genuine differences in job content and skill requirements? In particular, is internal wage structure best characterized as reflecting "objective long-term job values," management aims of controlling workers, or the ongoing effects of bargaining between firms, and workers and their organizations? 3) Does exclusive firm reliance upon internal promotion reflect enterprise-specific human capital, pressures from the workforce, or some other set of factors? Is the observed range of indeterminacy in wage structure better explained by enterprise-specific human capital and worker risk preferences, or by institutional rigidities?

In the case of the U.S. iron and steel industry, internal labor markets first became

prevalent in the latter nineteenth century.<sup>1</sup> At the time, unionization was largely confined to process workers who served as inside contractors, and directed crews with strategic responsibility for operations at bottlenecks in production activity. These process workers reluctantly changed their union rules and gave up traditional rights of interplant mobility in response to demands from lower-ranking crew members for internal promotion. Similarly, in Britain, internal promotion was decided upon at the founding convention of a less-exclusive rival of the contractors' union. In both countries, the demands of crew members for promotion rights were raised amid society-wide restiveness among less-skilled workers, and changes in iron and steel technology, industrial organization, and management methods which increased the vulnerability of exclusive craft unionism. These institutional influences on union policy appear to have outweighed whatever import enterprise-specific human capital had for iron and steel job structures. Furthermore, rather than unilaterally introducing internal labor markets, as in Stone's account, U.S. steel management only subsequently maintained internal promotion practices that had already been established under nineteenth-century trade unionism.

In the United States as well as Britain, the wage structure that developed by the turn of the century was shaped by the differential bargaining leverage of what was, by then, a somewhat larger group of unionized process workers. Unionized operatives, unlike non-union labor, were paid wages based on rates of tonnage output, a form of incentive payment which, in general, causes difficulties for managerial administration. Frequent rate cuts tend to destroy the efficacy of incentives. On the other hand, technical change may increase output, regardless of effort, and make work standards difficult to determine, creating an upward ratchet effect on incentive earnings. In iron and steel, these tendencies for "wage drift" were reinforced by union policy and bargaining power. In the

<sup>1</sup>For a more detailed discussion, see my 1983 study.

United States, the national union sought to maintain a standard tonnage rate policy. In Britain, and in U.S. departments where technical conditions were more dynamic and heterogeneous, tonnage rates were characteristically settled by shop management and small work groups, within a second tier of bargaining that had wide autonomy from national employers' associations and trade unions. The result of either bargaining arrangement was that tonnage rates proved downwardly rigid, despite rapid technological change, and tonnage wage earnings varied directly with plant productivity, as union behavior approximated that of discriminating monopoly. Consequently, in technologically dynamic departments that were union strongholds, wide occupational pay differentials developed.

In the United States, the wage structure established during the union era outlasted a series of strike defeats in 1892, 1901, and 1909 which all but destroyed union organization and bargaining influence. In 1910, despite preceding reductions in tonnage rates, wage spreads between top paid operatives and laborers were as wide as 6:1 in some production departments. By contrast, in blast furnaces, where neither contracting nor unionism had made significant inroads, and operatives had little control over the rate and quality of output, pay spreads were only 1.4:1. In general, departments with wider occupational pay differentials also had greater average pay—a rough indicator of expected career earnings on the different job ladders open to workers within the establishment. For example, in blast furnaces, average pay was only marginally greater than the going rate for laborers, while in bar mills average pay was 70 percent greater than the laborers' rate.

The differential incidence of union organization and productivity growth in the British and U.S. industries created disparities in their comparative wage structure. Although occupational wage differentials, in general, have been considerably wider in the United States than in Britain—a fact conventionally attributed to an historic scarcity of skilled labor in the United States—by 1910, in several

iron and steel production departments, British occupational pay differentials exceeded that in the United States. The relative position of iron and steel production departments by average pay also differed substantially between the two countries. Yet the two industries employed similar basic technology, and detailed job descriptions indicate their process workers performed much the same tasks.

Neither radical or neoclassical theory provides an adequate framework for comprehending this wage structure. Were job and pay structures, as radicals maintain, unrelated to technological differences in job content, and designed in accordance with managerial aims of controlling workers, pay differentials within departments should have been neither so narrow as to dampen promotion incentives, nor so wide as to add prohibitively to average wage costs. But this implication of radical theory is belied by the slight pay differentials and low average pay observed in blast furnaces, and the much greater differentials and average pay in other U.S. departments. By substantially affecting the average pay positions of career job ladders, the impact of bargaining on wage structure also overran the bounds of indeterminacy set by enterprise-specific human capital. The system of collective regulation within iron and steel internal labor markets neither replaced small group bargaining, nor yielded results that were rational by the efficiency standards of neoclassical theory. On the contrary, internal labor markets afforded greater leeway for an impact on wage structure of small group and national bargaining by erecting barriers to labor market competition. While steel firms paid a common regional entry wage for laborers, no upper constraint appears to have been binding on occupational wage structure, as senior process workers, contrary to Raimon, earned far more than skilled maintenance craftsmen with interindustry mobility.

In the United States, from 1910 to 1937, during the industry's nonunion era, steel firms observed a pattern of wage, as well as price leadership, and filled upper level job vacancies exclusively by internal promotion.

Occupational wage differentials were substantially compressed. On the other hand, establishment wage structures remained highly disparate, although published descriptions and the attitudes of the parties indicate job content was generally comparable throughout the industry. Upon the reestablishment of unionism in 1937-42, steel firms were deluged with wage inequity grievances alleging unequal pay for equal work, and subsequently negotiated an industrywide job evaluation and pay plan which drastically reduced wage dispersion within occupations. In Britain, on the other hand, no job evaluation plan was ever adopted, but after 1940 pay structure was substantially compressed through flat rate cost-of-living increases. Despite these intervening pay movements, in 1957, U.S. and British comparative wage structure displayed the same basic disparities as in 1910, albeit reduced in magnitude. The result of rigid internal promotion practices and ongoing informal bargaining appears to have been the preservation of certain basic aspects of customary wage structure, along with a latitude for occupational wage determination which continued to exceed the bounds of enterprise skill idiosyncrasy.

### III. Conclusions

In brief, iron and steel industry experience suggests the primary cause of internal labor markets is pressure collectively exerted by workers for employment security and advancement, with consequences which may include rigid internal promotion rules. While mutually supportive, these two points are logically distinct and have their own wider implications. In at least a number of other U.S. industries, unions appear to have been responsible for introducing greater employment stability and regulation. But internal promotion rules can prove rigid, whatever their origins.

In general, within U.S. unionized industries, however internal promotion rules originated, their rigidity is enforced by collective bargaining agreement. No comparable source of rigidity exists for nonunion firms, which may, and at times do, depart from conventional practice to fill upper-level

job vacancies by external hiring. Nonetheless, rigid promotion rules apparently prevail over wider terrain and have deeper roots than can be explained by trade union impact (for example, see Paul Osterman, 1983). Judging by steel industry experience, four main factors appear to explain why even nonunion firms often adhere invariably to internal promotion practices once they are established throughout an industry: the training and production requirements of modern business enterprise; implicit oligopolistic agreements; pressures from the workforce; and the demands of economic growth.

In general, the mass production technology employed by large business enterprise requires reliable access to input supplies (Alfred Chandler, 1977). Once established industrywide, internal labor markets assure firms of access to adequate supplies of trained workers. With interfirm mobility for industry-specific jobs generally foreclosed, external hiring which occurs in lieu of internal promotion reduces the career earnings prospects of the permanent workforce, and damages employee morale and the reputation of the enterprise in the labor market. There being no industry market for skilled labor, the enterprise is also unlikely to base production plans on marginal additions to its permanent workforce through external hiring of skilled workers, as it cannot assume the workers it wants can be recruited at a given wage.

Poaching of skilled labor is also likely to be viewed, much like price cutting, as a breach of an implicit oligopolistic agreement that redounds to the general detriment of employers. In turn, because external hiring undermines the monopoly position of enterprise workers, they are likely to view it as contrary to their collective interest. By increasing the risk of acute scarcity of skilled labor, and aggregate industry training requirements, growth can reinforce the effects of each of the factors previously mentioned.

By implication, internal promotion rules are more likely to be rigid for production workers within large firms and for workers prone to collective organization. This accords with the casual observation that, in the United States, external hiring is more com-

monly a live option in small firms, and for managerial, technical, craft, clerical, and service personnel.

The resulting latitude for wage structure appears to be constrained, to some extent, by worker risk preferences, as evidenced by the prevalence of a common entry level wage rate in iron and steel and in other industries. However, with bilateral monopoly prevailing between the firm and various groups in the workforce, there is no good reason to expect wage structure, overall, to conform closely to any simple optimizing model. Aside from the entry rate, the chief operative constraints on iron and steel wage structure impinged on expected career wage earnings and the average establishment wage. These findings suggest there may generally be wide bounds for institutional influence on wage structure.

#### REFERENCES

- Chandler, Alfred D. Jr., *The Visible Hand*, Cambridge: Harvard University Press, Belknap Press, 1977.
- Doeringer, Peter and Piore, Michael, *Internal Labor Markets and Manpower Analysis*, Lexington: D.C. Heath and Co., 1971, 13-40.
- Elbaum, Bernard, "The Making and Shaping of Job and Pay Structures in the Steel Industry," in Paul Osterman, ed., *Employment Practices within Large Firms*, Cambridge: MIT Press, 1983 forthcoming.
- Freeman, Richard, "Union Wage Practices and Wage Dispersion Within Establishments," *Industrial and Labor Relations Review*, October 1982, 36, 3-22.
- \_\_\_\_\_, "Unionism and the Dispersion of Wages," *Industrial and Labor Relations Review*, October 1980, 34, 3-23.
- Gordon, David, Edwards, Richard and Reich, Michael, *Segmented Work, Divided Workers*, Cambridge: Cambridge University Press, 1982.
- Osterman, Paul, "White Collar Internal Labor Markets," in Paul Osterman, ed., *Employment Practices in Large Firms*, Cambridge: MIT Press, 1983 forthcoming.
- Raimon, Robert L., "The Indeterminateness of Wages of Semi-Skilled Workers," *Industrial and Labor Relations Review*, January 1953, 6, 180-94.
- Stone, Katherine, "The Origins of Job Structures in the Steel Industry," *Review of Radical Political Economics*, Summer 1974, 6, 61-97.
- Williamson, Oliver, Wachter, Michael and Harris, Jeffrey, "Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange," *Bell Journal of Economics*, Spring 1975, 6, 250-77.

## RECENT ADVANCES IN THE THEORY OF INDUSTRIAL STRUCTURE

### Raising Rivals' Costs

By STEVEN C. SALOP AND DAVID T. SCHEFFMAN\*

Conduct that unreasonably excludes competitors from the marketplace is a concern of antitrust law. Predatory pricing doctrine focuses on conduct that lowers revenues. Alternatively, a firm can induce its rivals to exit the industry by raising their costs. Some nonprice predatory conduct can best be understood as action that raises competitors' costs.

To a predator, raising rivals' costs has obvious advantages over predatory pricing. It is better to compete against high-cost firms than low-cost ones. Thus, raising rivals' costs can be profitable even if the rival does not exit from the market. Nor is it necessary to sacrifice profits in the short run for "speculative and indeterminate" profits in the long run. A higher-cost rival quickly reduces output, allowing the predator to immediately raise price or market share. Third, unlike classical predatory pricing, cost-increasing strategies do not require a "deeper pocket" or superior access to financial resources. In contrast to pricing conduct, where the large predator loses money in the short run faster than its smaller "victim," it may be relatively inexpensive for a dominant firm to raise rivals' costs substantially. For example, a mandatory product standard may exclude rivals while being virtually costless to the predator.

These elements combine to make cost-increasing strategies more credible than

predatory pricing. Because these strategies do not require a sacrifice of profits in the short run, but allow profits to be increased immediately, the would-be predator has every incentive to carry out its threats. Moreover, unlike predatory pricing, cost-increasing strategies can often be made irreversible, and thus more credible.

Legal rules governing cost-increasing conduct should differ from predatory pricing standards. Price-cost comparisons alone are insufficient because such comparisons cannot distinguish price decreases from cost increases. Moreover, in some cases concerning conduct that raises rivals' costs, courts do not need to strike the difficult balance between short-run welfare gains and long-run losses. There is often no tradeoff. Cost increases generally raise prices, not lower them.

A variety of exclusionary practices can be characterized as conduct that raises rivals' costs. In the famous *Klor's* group boycott case, for example, the alleged predator Broadway-Hale may have induced a significant number of suppliers to refuse to provide needed inputs to Klor's. If these firms were the most efficient suppliers, a boycott could have raised Klor's costs and thus placed it at a competitive disadvantage. Had Klor's been a significant competitor in the market, retail prices could have been increased. Inducing suppliers to discriminate against rivals is a less extreme variant of the same conduct. Similarly, according to Oliver Williamson's 1968 analysis of the *Pennington* case, an industrywide wage contract raised the costs of the labor-intensive competitive fringe more than it raised the costs of the more capital-intensive dominant firms.

If there are scale economies or other entry barriers in retailing, exclusive dealing arrangements can raise small rivals' costs of distribution. As emphasized in the rent-seek-

\*Professor of economics, Georgetown University Law Center, and professor of economics and Director, Institute of Applied Economic Research, Concordia University, respectively. Some of the issues analyzed here were discussed in Salop (1981), and in our joint 1982 study. Judith Gelman has provided helpful comments. Financial support from the Bureau of Economics of the Federal Trade Commission is gratefully acknowledged. This paper does not reflect the views of the Commission, individual Commissioners, or staff.

ing literature, product standards and other government regulations can raise rivals' relative compliance costs. Advertising expenditures and *R&D* races can also be used to raise rivals' costs. For example, suppose that increased advertising expenditures initiated by the most efficient advertiser must be matched in effective intensity by less efficient rivals. An advertising strategy might be profitable even absent the demand-increasing effect of the advertising. Disadvantaging competitors can provide a benefit that exceeds its costs, if the strategy allows the dominant firm to increase price or market-share.

Though currently out of fashion with anti-trust enforcers, vertical price squeezes can be viewed as conduct to raise rivals' costs. Under appropriate conditions, a dominant firm finds backward integration to be a cost-effective way to raise downstream prices.<sup>1</sup> If the upstream merger partner has some market power, input price increases to downstream rivals (perhaps to a level *above* the monopoly price) will raise their costs, allowing the dominant firm to increase price or output. Upstream profits are sacrificed but downstream profits rise disproportionately.

The rest of this paper provides brief diagrammatic and formal analyses of these strategies. Our results are discussed intuitively; the technical analysis is taken up elsewhere. Three conditions are discussed: profitability to the dominant firm; injury to rivals; and consumer welfare losses. These conditions are then related to analogous concepts in the antitrust law of exclusionary practices.

Consider an industry consisting of a dominant firm and a competitive fringe. In such an industry, a lower-cost dominant firm acts as price leader. Competitive fringe firms follow by collectively setting some output  $y$  on the fringe supply curve  $S$ . Because each fringe

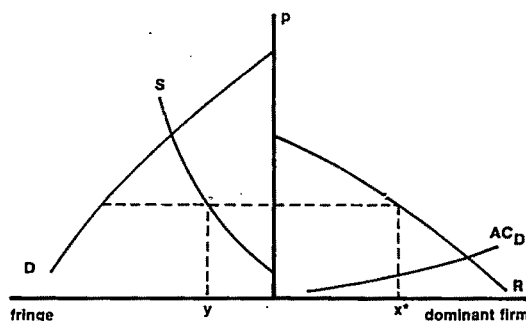


FIGURE 1

firm is small, it produces until price equals marginal cost. Indeed, for analytic simplicity, the supply curve is sometimes treated as if arising from a representative firm's marginal cost curve. At the equilibrium for such an industry, the dominant firm produces at the profit-maximizing point  $x^*$  on its residual demand curve  $R$ , as illustrated in Figure 1. The industry demand ( $D$ ) and the fringe supply (marginal cost) curves are shown in the left panel. The dominant firm's residual demand ( $R$ ) and average cost ( $AC_D$ ) curves are pictured in the right panel. Its profits are equal to  $(p - AC_D)x^*$ .

Suppose the dominant firm can also select a second strategy variable to which the fringe firms must react. A general way to view this strategy is to treat the firm as selecting a "technology" for producing output or revenue. Technologies differ in cost; each fringe firm reacts by choosing a technology itself. Particular strategy variables might include product quality or advertising expenditures. Another potential instrument is the demand for necessary inputs or, alternatively, the price offered for those inputs by the dominant firm. Labor, scarce natural resources, and patentable innovations are three inputs that have concerned antitrust commentators. (Williamson, M. Maloney et al., Richard Gilbert, and Janusz Ordover-Robert Willig.) Nonprice vertical restraints like exclusive dealing and territorial restraints can also be captured in this way, because they can affect the costs of distribution. The rent-seeking literature treats cases where a firm enters the political arena in order to inflict costly regu-

<sup>1</sup>In the limiting case of an upstream monopolist and downstream fixed proportions (and constant returns to scale) technology, it is well known that a vertical price squeeze is unnecessary. However, few industries satisfy this structure. In other cases, vertical price squeezes can be profitable under appropriate conditions.

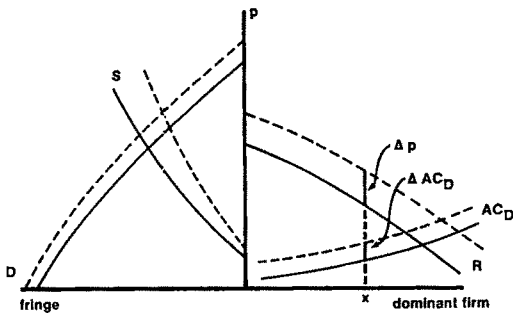


FIGURE 2

lations on its rivals, and possibly even itself (Maloney and R. McCormick).

The dominant firm's strategy may affect its own costs and market demand as well as the costs of its fringe competitors. As illustrated in Figure 2, a *sufficient* condition for a strategy to be profitable is for it to shift up the dominant firm's *residual demand curve* by more than it shifts up its *average cost curve* at the original output  $x^*$  (see equation (3) below). In this way, even if the dominant firm were to keep its output constant, the increased price-cost margin would raise its profits. Of course, the predator can generally increase its profits still further by adjusting its output.

Even if market demand is unaffected by the strategy, increases in marginal costs can reduce fringe firms' outputs and/or raise price, as illustrated in Figure 3. Sufficient increases in average costs can cause some fringe firms to exit the industry and others to forego entry. Thus, the concept of strategically erected entry barriers can be captured in this framework.

The shift in the residual demand curve depends on the elasticity of demand as well as the elasticity and shift of the fringe supply curve. The less elastic is consumer demand, the greater will be the increase in residual demand. This is because as demand elasticity falls, a given reduction in fringe supply causes a larger price rise (see equation (2) below). At the other extreme, if demand is perfectly elastic, residual demand does not increase at all.

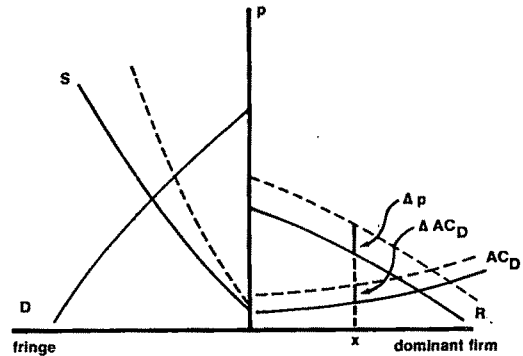


FIGURE 3

Suppose the fringe supply curve is treated as the marginal cost curve of a representative fringe firm. Under this interpretation, the dominant firm's residual demand curve shifts up according to the increase in the fringe's marginal costs, weighted by the elasticity of the market demand curve. As illustrated in Figure 3, evaluating the profitability of the strategy requires a comparison of this price rise to the increase in the average cost of the dominant firm. Thus, in effect one must compare the effect on the *average cost* of the dominant firm relative to the *marginal cost* of the fringe, weighted by the demand elasticity.

Formally, when demand is unaffected by the strategy, the dominant firm's optimization problem is given as follows:

$$(1) \quad \text{Max } px - C(x, \alpha)$$

$$\text{subject to } x = D(p) - S(p, \alpha); \alpha \geq 0,$$

where positive adoption of a strategy is formulated as choosing  $\alpha > 0$ , the dominant firm's costs  $C(x, \alpha)$  are assumed to depend on its output  $x$  and the strategy  $\alpha$ , and its residual demand consists of industry demand  $D(p)$  less fringe supply  $S(p, \alpha)$ .

In solving this problem, a *sufficient* condition for a strategy  $\alpha > 0$  to be chosen is given as follows:

$$(2) \quad 1/(1 + \sigma\epsilon/(1 - \sigma)\epsilon_s) > \Delta AC_D / \Delta MC_F,$$

where  $\epsilon$  and  $\epsilon_s$  represent the elasticity of industry demand and fringe supply, respectively,  $\sigma$  is the market share of the dominant firm, and  $\Delta AC_D$  and  $\Delta MC_F$ , respectively, represent the strategy-induced changes in the dominant firm's average cost ( $C_a/x$ ) and the representative fringe firm's marginal cost (which equals  $-S_a/S_p$ ). Differentiating the constraint in equation (1), substituting into equation (2) and rewriting, the sufficient condition becomes

$$(3) \quad \partial p / \partial \alpha|_{x^*} > \Delta AC_D.$$

We have so far discussed only the effect of the strategy on the profitability of the dominant firm. The profitability of the fringe and consumer welfare will also be affected.

In principle, fringe output and profits may rise or fall. There are two separable effects on the fringe. First, its costs rise, lowering fringe profits. Second, because the dominant firm chooses a new output price, the profits of the fringe are further affected. These two effects generally work in opposite directions, of course, because increases in marginal cost cause price increases. For example, consider the limiting case of perfectly inelastic demand. If the fringe output is held constant, its profits fall if the strategy raises its average cost by more than the increase in fringe marginal cost. This is because, holding outputs constant, price rises by the increase in fringe marginal cost. A reduction in the fringe's output reinforces this effect whereas production increases offset the effect of the reduced profit margin.

If competitors' profits are not reduced, the strategy will obviously fail to achieve an exclusionary goal. However, if the industry is protected by entry barriers, strategies that increase the costs of fringe firms and dominant firm *equally* can still raise industry profits. In particular, if marginal costs rise by more than average costs and if demand is sufficiently inelastic, the cost increases will have a *supra*-passing-on effect, raising price by more than the increase in average cost.

Consumers' surplus is also affected by these cost-increasing strategies. Again, cost

increases tend to cause price increases, which are welfare reducing. However, there may be cases in which demand and supply elasticities are increased sufficiently to cause price to fall enough to offset the welfare losses from the higher costs. In addition, in that demand (i.e. marginal consumers' surplus) is increased, consumers' surplus may rise even at a higher price. Similar results obtain for measures of aggregate economic welfare (consumers' surplus plus profits). For example, a strategy that does not raise demand, yet raises cost and price, surely lowers aggregate welfare. If demand rises, however, price, cost, and demand increases must be balanced.

For antitrust analysis, exclusionary strategies may be characterized by three conditions—profitability to the dominant firm; competitor injury; consumer welfare reduction—and their sum, the allocational efficiency (or aggregate welfare) effect. One formulation of the attempt to monopolize offense—unreasonable conduct undertaken with specific intent to monopolize that has a dangerous probability of success—can be interpreted in terms of these conditions. Long-run profitability to the dominant firm is an obvious element of intent to monopolize. Competitor injury is necessary for the conduct to have a dangerous probability of success. A strategy that reduces consumer welfare or allocational efficiency might well satisfy the unreasonableness prong of the offense.

## REFERENCES

- Gilbert, Richard, "Patents, Sleeping Patents and Entry Deterrence," in S. Salop, ed., *Strategy, Predation and Antitrust Analysis*, Federal Trade Commission Report, 1981.
- Maloney, M. and McCormick, R., "A Positive Theory of Environmental Quality Regulation," *Journal of Law and Economics*, April 1982, 25, 99–124.
- \_\_\_\_\_, \_\_\_\_\_, and Tollison, R., "Achieving Cartel Profits Through Unionization," *Southern Economic Journal*, October 1979, 42, 628–34.

Ordover, Janusz and Willig, Robert, "An Economic Definition of Predation: Pricing and Product Innovation," *Yale Law Review*, November 1981, 91, 8-53.

Salop, S. C. "Introduction," in his *Strategy, Predation and Antitrust Analysis*, Federal Trade Commission Report, 1981.

\_\_\_\_\_ and Scheffman, D. T., "Non-Price Predation by a Dominant Firm," unpublished paper, 1982.

Williamson, Oliver, "Wage Rates as a Barrier to Entry: The Pennington Case," *Quarterly Journal of Economics*, February 1968, 85, 85-116.

# The Welfare Effects of Intermittent Interruptions of Trade

By GLENN C. LOURY\*

Consider the problem of a buyer purchasing a (costlessly) storable good at a fixed price under threat of interruption of trade. Suppose there are two distinct and exhaustive regimes: either one can buy as much as desired at the going price, or one can purchase nothing at all and must rely on accumulated stocks for consumption. The market moves randomly over time between these two states, which will be referred to here as "on" and "off," respectively. Given the stochastic process governing the evolution of market regimes, the price faced when the market is on, and the intertemporal preferences of the buyer, one may inquire as to the optimal acquisition rate for storage and consumption when the market is on, and the optimal usage of accumulated stocks when the market is off. Solutions for these problems could then be used to ascertain the degree to which buyer (and seller) welfare is affected by the probabilistic curtailment of trading opportunities, and the willingness to pay of the buyer for greater security of supply. As well, the possibility of inferring from purchase and storage behavior the buyer's perception of the likelihood and duration of trade interruptions may be explored.

This paper presents a general solution to this problem for stationary environments. That is, assuming intertemporal preferences to be representable by the expected present value of a uniformly discounted, time-independent flow utility function, and taking the stochastic process governing change in regimes to be stationary Markov, optimal behavior and associated expected discounted utility are derived. The method employed adapts to this context the dynamic programming arguments of my 1981 paper, and is similar in structure to that employed in the

analysis of resource depletion with technological uncertainty in Partha Dasgupta and Geoffrey Heal (1974) and Dasgupta and Joseph Stiglitz (1981).

There are many industrial markets in which these considerations are important. Recent events in the world oil market have stimulated much interest in the study of strategic mineral reserves (see, for example, Thomas Teisberg, 1981). In markets where demand fluctuates randomly and prices are inflexible (for example, the industrial market for natural gas), the possibility of stochastic rationing has important implications for firm behavior (see Dennis Carleton, 1978). The prospect of strike upstream, or lag in the delivery of goods on order has significant implications for inventory holdings in some industries. The framework of this paper may even be adapted to the study of labor supply and savings behavior over the life cycle for a worker facing intermittent and uncertain unemployment.

## I. The Problem

Let  $\{\tilde{x}_t\}$  be a two-state, continuous time stochastic process with the interpretation that  $\tilde{x}_t = 0$  when the market is off, while  $\tilde{x}_t = 1$  when the market is on. Assume  $\{\tilde{x}_t\}$  is stationary Markov. Denote by  $\tilde{\tau}_i$ ,  $i = 0, 1$ , the elapsed time until leaving state  $i$ , given that  $\{\tilde{x}_t\}$  starts in state  $i$ . Then  $\tilde{\tau}_i$  is exponentially distributed with parameter  $h_i$ , and the process  $\{\tilde{x}_t\}$  is completely characterized by the two instantaneous "exit" probabilities  $(h_0, h_1)$ . This follows because the Markovian assumption implies that the likelihood of leaving state  $i$  in the next instant after date  $t$  is independent of the history of the process, while the stationarity assumption implies that likelihood, given the history, is independent of the date  $t$ .

Suppose that when the market is on, the buyer can exchange on fixed terms a domestically produced product for the good

\*Harvard University. I thank Clas Bergström and Mats Persson for helpful discussion. This paper extends ideas developed in our joint work (1982).

whose supply may be interrupted. The buyer carries at any date a nonnegative stock of accumulated supplies of this good, which may be drawn down or added to as he chooses. Denote by  $z$  the (nonnegative) quantity of domestic product supplied, by  $q$  the (nonnegative) amount of external good consumed, by  $w$  the fixed terms of trade, and by  $S$  the stock of external supplies on hand at any date. Then the stock on hand evolves according to the equation

$$(1) \quad \dot{S}_t = wz_t x_t - q_t.$$

Let  $u(q, z)$  be the buyer's instantaneous payoff function, assumed strictly concave, increasing in  $q$ , and decreasing in  $z$ . Denote the buyer's rate of discount by  $r$ . For future reference, define the function  $u^*$  as follows:

$$(2) \quad u^*(x, p) = \text{Max}_{q, z} \{u(q, z) + p(wzx - q)\}.$$

Of course, with  $x = 0$ , the solution involves setting  $z = 0$ . Finally, let  $(q^*, z^*)$  solve

$$\text{Max}_{q, z} \{u(q, z) | wz - q = 0\},$$

and define

$$(3) \quad p^* \equiv \partial u(q^*, z^*) / \partial q.$$

The buyer's problem is to choose a path for  $(q, z)$  to maximize expected discounted utility subject to initial conditions  $(\bar{x}, \bar{S})$ , the differential equation (1), and the probability law for  $\{\tilde{x}_t\}$ . Assume the existence of a solution for this problem. Define the value function  $V$  such that  $V(x, \bar{S})$  gives the maximal expected discounted utility from initial conditions  $(\bar{x}, \bar{S})$ . Defining  $y = 1 - x$ , it is clear that  $V(\cdot, \cdot)$  must satisfy the system of Bellman Equations:

$$V(x, \bar{S}) = \text{Max} \left\{ \int_0^\infty h_x e^{-h_x t} \left\{ e^{-rt} V(y, S_t) + \int_0^t e^{-rs} u(q_s, z_s) ds \right\} dt, \right.$$

subject to equation (1),

$$S_0 = \bar{S}, \text{ and nonnegativity} \}.$$

This may in turn be simplified to yield the basic relation:

$$(4) \quad V(x, \bar{S}) = \text{Max} \left\{ \int_0^\infty e^{-(r+h_x)t} [h_x V(y, S_t) + u(q_t, z_t)] dt, x + y = 1, \right.$$

subject to equation (1),

$$S_0 = \bar{S}, \text{ and nonnegativity} \}.$$

We proceed heuristically by treating (4) as an ordinary optimal control problem, applying the maximum principle. First, suppose  $x = 1$ . The current value Hamiltonian may be written

$$H(1, p_t, S_t) = u^*(1, p_t) + h_1 V(0, S_t),$$

where "1" denotes the regime,  $p_t$  is the co-state variable, and  $u^*(\cdot)$  is defined above. Then, along the optimal path,  $(p_t, S_t)$  must satisfy

$$(5) \quad \dot{p}_t = (r + h_1) p_t - h_1 \partial V(0, S_t) / \partial S;$$

$$\dot{S}_t = \partial u^*(1, p_t) / \partial p.$$

A standard phase diagram analysis shows that (5) has a unique equilibrium point which is a saddle point.<sup>1</sup> As usual, given  $S_0 = \bar{S}$ , the initial price  $p_0$  must be chosen so that the solution path for (5) from those initial conditions approaches this equilibrium. It is clear that at this equilibrium point,  $(p, S)$  satisfies the equations  $(r + h_1)p = h_1 \partial V(0, S) / \partial S$  and  $0 = \partial u^*(1, p) / \partial p$ . But the second equation can only hold if  $p = p^*$ , as defined by (3). Thus the equilibrium stock  $S^*$  is defined by

$$(6) \quad p^*(\partial V(0, S^*) / \partial S) h_1 / (r + h_1).$$

Now, in a world of no interruptions in trade  $p^*$  is the marginal utility of  $q$  at the

<sup>1</sup>This depends on convexity of  $u^*$  in  $p$ , and concavity of  $V$  in  $S$ . The former follows from definitions. The latter can be shown to follow from the assumed concavity of  $U(q, z)$ .

optimal consumption pair  $(q^*, z^*)$ . Equation (6) states that when such interruptions are possible, stocks should be accumulated until their shadow value exceeds this marginal utility by the factor  $(r + h_1)/h_1$ . Once this level is reached consumption continues at  $(q^*, z^*)$  until trade is interrupted. (Though, if  $S_0 \neq S^*$ , the latter is approached only asymptotically.) It follows that the value of best behavior starting from  $(1, S^*)$  is given by

$$(7) \quad V(1, S^*) = V(0, S^*)h_1/(r + h_1) \\ + u^*(1, p^*)/(r + h_1).$$

Now consider the maximization problem, (4), with  $x = 0$ . It is of course, equivalent to a depletable resource problem with an uncertain introduction of new "technology" which leaves the remaining stock valued at  $V(1, S)$ . Application of dynamic programming arguments may be used to show that

$$V(0, S) = V(1, S)h_0/(r + h_0) \\ + \text{Max}_q \{u(q, 0) - q \partial V(0, S)/\partial S\}/(r + h_0).$$

Making use of (6), one has then that when the optimal stock is in hand, welfare immediately after an interruption satisfies

$$(8) \quad V(0, S^*) = V(1, S^*)h_0/(r + h_0) \\ + u^*(0, p^*(r + h_1)/h_1)/(r + h_0).$$

Equations (7) and (8) represent two linear equations in the unknowns  $V(0, S^*)$  and  $V(1, S^*)$ . They may be readily solved to express these values in terms of the data of the problem (i.e., preferences and probabilities.) Solving for  $V(x, S^*)$  yields

$$(9) \quad V(1, S^*) = \frac{r + h_0}{r + h_0 + h_1} \frac{u^*(1, p^*)}{r} \\ + \frac{h_1}{r + h_0 + h_1} \left( \frac{u^*(0, p^*(r + h_1)/h_1)}{r} \right);$$

$$V(0, S^*) = \frac{h_0}{r + h_0 + h_1} \frac{u^*(1, p^*)}{r} \\ + \frac{r + h_1}{r + h_0 + h_1} \left( \frac{u^*(0, p^*(r + h_1)/h_1)}{r} \right).$$

Thus, with the optimal stock in hand, welfare is the weighted average of two perpetual flows, denoted  $u_1^*$  and  $u_0^*$  for convenience. The numbers  $u_i^*$ ,  $i = 0, 1$ , can be interpreted as indirect utilities. Condition (9) reveals that facing the random access to market at fixed terms of trade is equivalent from a welfare point of view to facing continuous access to the market, but at random terms of trade, once the optimal stock is in hand. The problem with this observation is that if initial stock is less than  $S^*$ , then with probability one the optimal stock is never reached. One has in effect convex "costs of adjusting" the stock to the optimal level. If, however, the marginal rate of substitution between  $q$  and  $z$  is assumed independent of  $z$  (i.e., constant marginal utility of  $z$ ), then "adjustment costs" are linear and the optimal stock is approached in a "bang-bang" fashion. One is then always at the optimal stock when the market is on, and (9) becomes a powerful tool.

## II. A Special Case

Consider then the case where  $u(q, z) = \bar{u}(q) - z$ . One can think of this as a case of "constant marginal utility of income." Then  $\bar{u}(q)$  represents the monetary value of having the flow  $q$  of potentially interruptible supplies available for use at any date,  $z$  is the expenditure on market supplies when the market is on, and  $w$  is the inverse of the price of these supplies. This assumption seems strong if the buyer is thought of as a consumer choosing labor supply and consumption subject to a budget constraint, but reasonable if the buyer is a firm or country with access to a perfect capital market, seeking to maximize the expected present value of profits. The point of this assumption is that, with  $u(q, z)$  so defined,  $u^*(x, p)$  is independent of  $x$  when  $p = 1/w$ . Moreover, under this

assumption,  $p^* = 1/w$ . Thus, in this case,  $u_1^* = u^*(0, 1/w)$  and  $u_0^* = u^*(0, (r + h_1)/h_1 w)$ . To simplify notation, I will drop the first argument of  $u^*$  for the rest of this paper, and denote  $1/w$  as  $p$ . Our constant marginal utility assumption then implies

$$(9') \quad V(1, S^*) = \frac{r + h_0}{r + h_0 + h_1} \left( \frac{u^*(p)}{r} \right) + \frac{h_1}{r + h_0 + h_1} \left( \frac{u^*(p(r + h_1)/h_1)}{r} \right).$$

$$V(0, S^*) = \frac{h_0}{r + h_0 + h_1} \left( \frac{u^*(p)}{r} \right) + \frac{r + h_1}{r + h_0 + h_1} \left( \frac{u^*(p(r + h_1)/h_1)}{r} \right)$$

Thus, given that the optimal inventory is in hand, it is as if the buyer faced, in perpetuity, continuous access at prices randomly fluctuating between  $p (= 1/w)$  and  $p(r + h_1)/h_1$ , with the probabilities as given in (9').

The utility of this observation is illustrated by an application of Jensen's Inequality to (9'), recalling the convexity of  $u^*(p)$ . One may conclude that

$$(10) \quad V(1, S^*) \geq u^*(2p)/r.$$

That is, once the optimal inventory has been acquired, the threat of interruption can never make the buyer worse off than a doubling of the price with certain access to the market. Suppose the buyer, for fixed outlay  $K$ , can acquire the capacity to produce the embargoed good at constant marginal cost  $c > p$ . Since the optimal stock is approached immediately when the market is on, assuming constant marginal utility of  $z$ , the buyer will make a fixed expenditure of  $pS^*$  to acquire the optimal stock. Investing in the "synfuels" project dominates holding the stockpile then if and only if

$$(11) \quad r(K - pS^*) < h_1(r + h_0 + h_1)^{-1} (u^*(c) - u^*(p(r + h_1)/h_1)).$$

Thus, if the back-up technology has marginal

cost in excess of  $p(1 + r/h_1)$ , its development costs must be less than the market value of the optimal inventory for it to be a worthwhile option.

The optimal stock,  $S^*$ , can be explicitly derived in this case as well. Since (with no constraint on the rate of purchase at fixed price  $p$ ) the stock is adjusted immediately to its optimal level when the market is "on," it follows that  $V(1, S) = p(S - S^*) + V(1, S^*)$ . Substitution into equation (4) and integration by parts reveals that  $V(0, S)$  must satisfy

$$(12) \quad V(0, S) = \frac{h_0}{r + h_0} V(1, S^*) + \text{Max} \left\{ \int_0^\infty e^{-(r+h_0)t} (\bar{u}(q_t) - q_t p h_0 / (r + h_0)) dt, \right. \\ \left. \text{subject to } \int_0^\infty q_t dt \leq S \right\}.$$

The second term above is simply the value of an optimal resource depletion problem where initial stock is  $S$ , discount rate is  $r + h_0$ , and "extraction cost" at the margin is  $p h_0 / (r + h_0)$ . Since the marginal value of stocks, when the market is on, is constant at  $p$ , the opportunity cost of consuming some of the stock when the market is off is the expected present value of  $p$  at the embargo termination date:  $E_{\tau_0}(p \cdot e^{-r\tau_0}) = p h_0 / (r + h_0)$ . Thus, during an embargo one consumes the available stock as if it costs  $p h_0 / (r + h_0)$  per unit to procure. Now the solution for this maximization problem has the property that the marginal utility of  $q$  net of "extraction cost" grows exponentially at the interest rate  $r + h_0$ , starting from a value just equal to the shadow price on the stock constraint. This latter is the same thing (from equation (12) and the envelope theorem) as  $\partial V(0, S) / \partial S$ . Recalling that, from the definitions,  $\bar{u}'(-\partial u^*(p) / \partial p) \equiv p$ , and that  $\partial V(0, S^*) / \partial S = p(r + h_1)/h_1$ , it follows that the optimal stock is given explicitly:

$$(13) \quad S^* = - \int_0^\infty \left( \partial u^*(h_0 p / (r + h_0) + e^{(r+h_0)t} p(r + h_1) / h_1 - h_0 / (r + h_0)) \right) / \partial p dt.$$

Equation (13) shows that observation of the inventory holdings of the buyer alone would be insufficient to separately identify the parameters  $h_0$  and  $h_1$ , although observing the initial "price" and its rate of change during an embargo would permit these effects to be determined.

### III. The Profitability of Trade Interruption

Since the buyer can store the good against the day when access to the market is lost, he can limit the damage of random interruptions, even when  $\bar{u}'(0) = \infty$ . This was noted in inequality (10). On the other hand, purchases for storage increase the seller's revenues during the period when the market is on. If demand is sufficiently inelastic and/or storage costs (in this case limited to interest costs of carrying stocks) sufficiently low, then this increase in purchases for precautionary holdings may enhance the seller's revenues by more than the expected loss in sales incurred during trade interruptions. However, with inelastic demand seller's revenues may also be increased by simply raising the price. This leads to the general question of whether any policy of the seller which involves the use of an embargo threat can be dominated by a policy without trade interruption, but with a different price. Since, by (9') probabilistically curtailing buying opportunities has welfare effects similar to the random increasing of price, intuition suggest such a strategy could be bettered by simply charging the profit-maximizing price all of the time. However, this intuition is false.

These issues can be investigated using the model of the preceding section. Assume  $h_0$ ,  $h_1$ , and  $p$  to be fixed. Let  $\Pi^*$  denote the EPDV of sales revenue given that buyer's initial stock is zero. Then this quantity is given by the following recursive relation:

$$\Pi^* = pS^* + E_1 \left\{ \int_0^{\tilde{\tau}_1} e^{-r't} p q^*(p) dt + e_1^{-r\tilde{\tau}_1} E_0 \{ e^{r\tilde{\tau}_0} (\Pi^* - pS\tilde{\tau}_0) \} \right\},$$

where  $E_i$  is the expectation operator under

the distribution of  $\tilde{\tau}_i$ ,  $i = 0, 1$ ,  $q^*(p) = -\partial u^*(p)/\partial p$  is the flow demand function, and  $S_i$  is the stock on hand when the embargo has duration " $t$ ." Carrying out the implied integration above, solving for  $\Pi^*$ , making use of (13) and manipulating yields

$$(14) \quad \Pi^* = \frac{r + h_0}{r + h_0 + h_1} \{ p q^*(p) / r + E_0 \{ p (S^* - S_{\tilde{\tau}_0}) e^{-r\tilde{\tau}_0} \} + \frac{h_1}{r + h_0 + h_1} \left\{ \int_0^\infty e^{-(r+h_0)t} (r + h_0) \times (p_t q^*(p_t) / r) dt \right\} \},$$

where  $p_t$  is the argument of  $\partial u^*(\cdot)/\partial p$  in (13), the shadow price of the good  $t$  units of time into an embargo. Equation (14) shows that the seller's welfare as measured by  $\Pi^*$  is the weighted average of two terms, with the same weights as appeared in the expression for buyer's welfare  $V(1, S^*)$ , given in (9'). The terms themselves may be readily interpreted. The first term is the present value of the perpetual revenue flow  $p q^*(p)$  plus the expected present value of purchases to replenish stocks at the termination of an embargo, given that an embargo has just begun. The second term is the average across time of the revenues which would accrue if the shadow price  $p_t$  were charged in perpetuity.

It is clear from (14) that periodic interruption of trade can be a profitable activity for the seller. With unitary elasticity of demand, for example,  $p_t q^*(p_t) = p q^*(p)$  for all  $t$ , so  $\Pi^* > p q^*(p) / r$ . Moreover, if demand is inelastic but the seller is not free to raise price, then he may nonetheless enjoy the effect on profits of the implicit price increase which intermittent interruption implies. When price is flexible but the seller desires to inflict a given welfare loss on the buyer while minimizing the loss of his own revenues, the use of an embargo strategy will generally be desirable.

#### IV. Conclusion

Several cautionary observations are in order. The preceding discussion has assumed the possibility of the seller effectively curtailing the buyer's access to the market. When there are several noncooperating sellers, this may be problematic. The results implying the possible profitability of an embargo strategy depend upon the assumed instantaneous adjustment of the buyer's buffer stocks to their optimal level. In the general case, stock adjustment occurs gradually, and these results need not hold. The "stationarity" assumption implies that the buyer has been observing the process over a period of time sufficiently long to have "learned" the relevant parameter values  $h_0$  and  $h_1$ . Buyer behavior during the transition to this "steady state" is not discussed here. Finally, also omitted from this discussion is consideration of the strategic interplay between buyer and seller, in which the "credibility" of seller's threats to interrupt trade, and the "deterrent effect" of buyer's stock-piling activity could be analyzed. Despite these limitations, the foregoing provides a simplified framework for the formal analysis of some important issues

which arise in the study of markets where trade interruptions periodically occur.

#### REFERENCES

- Bergström, Clas, Loury, Glenn and Persson, Mats, "Embargo Threats and the Management of Emergency Reserves," mimeo., Stockholm School of Economics, 1982.
- Carleton, Dennis, "Market Behavior with Demand Uncertainty and Price Inflexibility," *American Economic Review*, September 1978, 68, 571-87.
- Dasgupta, Partha and Heal, Geoffrey, "The Optimal Depletion of Exhaustible Resources," *Review of Economic Studies*, Symposium, 1974, 3-28.
- \_\_\_\_\_ and Stiglitz, Joseph, "Resource Depletion under Technological Uncertainty," *Econometrica*, January 1981, 49, 85-104.
- Loury, Glenn, "Pricing an Exhaustible Stock Subject to Stochastic Discovery," paper presented at Summer Meetings of the Econometric Society, La Jolla, June 1981.
- Teisberg, Thomas J., "A Dynamic Programming Model of the U.S. Petroleum Reserve," *Bell Journal of Economics*, Autumn 1981, 12, 526-46.

# Information, Competition, and Markets

By BARRY J. NALEBUFF AND JOSEPH E. STIGLITZ\*

One of the dominant characteristics of modern capitalist economies is the important role played by competition: not the peculiar static form of pure price competition embodied in the Arrow-Debreu model, but rather a dynamic competition, more akin to the kind of competition represented by sports contests and other races (including patent races).

In recent years, there have been several attempts to explain why firms often base the pay of their workers and managers on relative performance. (See, for example, Edward Lazear and Sherwin Rosen, 1981.) Such compensation schemes become desirable when three conditions are satisfied: (a) The input (effort) of workers (managers) must not be directly observable, at least without cost. Thus firms must either expend resources to monitor inputs or devise reward structures in which compensation is a function of variables (such as output or profits) which are themselves functions of inputs but are less costly to observe. (b) The relationship between input and output must be stochastic, so that by observing output, one cannot perfectly infer what the input was. (c) Finally, the stochastic disturbances which affect the relationship between input and output of different firms must be correlated. By looking at the performance of one worker relative to that of others, one can make better inferences about his effort than one can make without using this information.

Not only can competition provide a basis of comparison, which enables the design of reward structures that can simultaneously provide a high level of incentives with relatively low level risk; but compensation schemes based on relative performance have the further advantage of automatically ad-

justing incentives to changes in the economic environment. (We refer to this as "built-in flexibility.") In a first best world, with perfect information concerning the nature of the technology (but where it is still costly to monitor individuals' activities), the compensation scheme would vary from time to time as the environment changed. Such changes in the compensation scheme are costly to implement and the information required to do so is seldom available. When a task is easier, the individual's rewards for performing the task should be reduced. If pay is based on *relative* performance, although all individuals perform better (when they exert the same level of effort), their compensation is automatically adjusted. Thus, teachers frequently grade on the curve and a significant fraction of the pay of successful salesmen often consists of bonuses based on relative performance.

## I. Competition and Compensation

To see more clearly exactly how competition can provide the basis of the design of a better compensation scheme, we consider a simple example. Assume the government wishes to develop a bomber. Neither the government nor the potential developers know how much it should cost to build the bomber. With a fixed-fee contract (where the amount received by the developer is independent of his costs), the contractor will require a large risk premium to compensate for the large risk he must bear. The government can reduce the risk by sharing in the costs; but to the extent that it does this, it also reduces the contractor's incentives to save on costs. There is an alternative contractual arrangement that may be superior. Assume there are a number of potential contractors of the same ability, so that the costs faced by one firm (at any particular level of effort) will be identical to those faced by another firm. Assume the government lets out two contracts; it promises to pay firm *A* a fixed amount *plus*

\*Harvard University and Princeton University, respectively. We are indebted to Steve Salop, Sherwin Rosen, Ed Lazear, Felix Fitzroy, Joe Farrell, and Oliver Hart for helpful conversations. Financial support of the National Science Foundation is gratefully acknowledged.

whatever it costs firm  $B$  to produce the bomber; and conversely for firm  $B$ . Since their costs are perfectly correlated, this scheme eliminates all risk. At the same time, the scheme has perfect incentives: if firm  $A$  can, by exerting extra effort, reduce costs it gets to keep the savings. Having two separate firms provides us with information which simply would not be available otherwise. This information allows the implementation of a compensation scheme with lower risk and better incentives. There is a cost: the government has had to pay for duplicate research expenditure, but this may still be less expensive than the alternative.<sup>1</sup>

### A. General Theorems on the Optimality of Contests and Relative Performance Schemes

A natural question to raise at this point is, are there circumstances in which an appropriately designed contest or relative performance scheme can attain a first best allocation, that is, the same level of effort in every state as would obtain if information about the state were freely available and effort were costless to monitor, without the worker needing to bear any risk.

We consider a general structure in which the individual's output  $Q_i$  (assumed to be observable) is a function of a common random variable  $\theta$ , his effort  $\mu_i$ , and an idiosyncratic random variable  $\varepsilon_i$ ; for simplicity, we assume a linear relationship of the form:<sup>2</sup>

$$(1) \quad Q_i = \mu_i \theta + \varepsilon_i; E\varepsilon_i = E\varepsilon_i \varepsilon_j = E\varepsilon_j \theta = 0.$$

The worker observes  $\theta$  before he decides on effort; but the manager-owner of the firm

<sup>1</sup>With only two firms, there is the further danger of collusion. As the number of firms increases, the likelihood of collusion may decrease; but the excessive waste from duplication increases. In other contexts, however, such as natural monopolies with average cost curves only slightly declining, the loss in efficiency from having several competitors may be slight.

<sup>2</sup>What is crucial about this specification is that the common random variable affects the marginal productivity of labor; as a result, in the first best allocation, effort will change from state to state. Since one of the issues with which we are concerned is the extent to which compensation schemes provide the appropriate incentives to change the level of effort in response to changes in circumstances, it is important that the model analyzed have this feature.

can neither observe  $\theta$ ,  $\mu_i$ , nor  $\varepsilon_i$ ; he can only observe the output of each of his workers. We now show that if the distribution of  $\varepsilon_i$  is compact (as a normalization, we assume it ranges from  $-1$  to  $+1$ ), then with only two workers, it is possible to design an incentive structure which attains the first best. (This result is stronger than that of the earlier example, since the technology of the two individuals is not perfectly correlated.)

We assume, for simplicity, that workers have additively separable utility functions of the form

$$(2) \quad U(Y) - V(\mu),$$

where  $Y$  is income. With perfect information, a first best allocation requires  $Y = \bar{Y}$  (individuals obtain perfect insurance) and

$$(3) \quad \theta U(\bar{Y}) = V'(\mu).$$

(The marginal rate of substitution between leisure and goods equal the marginal rate of transformation.) The solution to (3) we refer to as the optimal level of effort, and denote by  $\mu^*(\theta)$ . Assume that after the individual has observed  $\theta$ , but before he has allocated his effort (and, in particular, before the output has been produced), the individual is asked to announce a "goal." He is told if he comes within one unit of making his goal, he will have a given income; if he fails to come within a unit of meeting his goal, he receives minus infinity (or a suitably large punishment). The higher the goal he sets, the higher the income he will receive, provided he attains it. Finally, he is told that his pay will depend on the announcement of others as well; they observe, of course, exactly the same common random variable that he does. We now show that there exists a compensation structure of the form indicated which provides perfect incentives and eliminates all risk.

Consider the compensation scheme which pays the  $i$ th individual

$$(4) \quad Y^i = \frac{\phi(\hat{\theta}^i)}{U'(\bar{Y})} - \frac{\phi(\bar{\theta}^{-i})}{U'(\bar{Y})} + \bar{Y}$$

if  $Q_i \geq \mu^*(\hat{\theta}^i)\hat{\theta}^i - 1$ ,  
 $-\infty$  otherwise;

where  $\hat{\theta}^i$  is the  $i$ th individual's announcement of  $\theta$ ,  $\bar{\theta}^{-i}$  is the average of the announcements of other individuals. If each contestant tells the truth, the individual faces no risk, since  $\phi(\hat{\theta}^i) = \phi(\bar{\theta}^{-i})$ . Moreover, the individual will, if he announces  $\theta$ , always make his target; he will choose his level of effort so that in the worst event ( $\varepsilon = -1$ ), he just makes it. The return to announcing a higher value of  $\theta$  is then

$$(5) \quad U'(\bar{Y})\phi'/U'(\bar{Y}) - V'd\mu/d\hat{\theta}.$$

To guarantee meeting the quota,

$$(6) \quad \mu^*(\hat{\theta})\hat{\theta} = \mu\theta,$$

$$(7) \quad d\mu/d\hat{\theta} = (\hat{\theta}/\theta)(d\mu^*/d\hat{\theta}) + \mu^*(\hat{\theta})/\theta.$$

Hence if

$$(8) \quad \phi' = V'[d\mu^*/d\theta + \mu^*(\theta)/\theta],$$

equation (5) will equal zero when  $\hat{\theta} = \theta$ . By integrating (8), we obtain a  $\phi$  function which, when used in the compensation scheme (4), provides perfect incentives and eliminates all risks.

Though the use of targets in conjunction with a relative performance scheme can, under some idealized conditions, provide the basis of an extremely effective compensation system, it has its limitations: if, for instance, individuals obtain information about their idiosyncratic random variable at different times, or if the support of the idiosyncratic random variable differs for different individuals or is not known by the employer. (If individuals know  $\varepsilon$  before the announcement occurs or if the announcement must be made before  $\theta$  is known, the announcement conveys no information additional to that which is conveyed by observing  $Q$ .)

### B. Further Results on Contests

In our forthcoming article, we investigate the design of compensation schemes which do not employ announcements. We have shown how contests can be designed to provide the first best level of effort in every state of nature. In general, such contests do impose a greater risk on the individual than he would bear if monitoring effort were costless. There are two conditions under which con-

tests may attain a first best outcome: (i) if the agents are risk neutral; or (ii) when there are a large number of contestants. Although generalized relative performance schemes include, as special cases, individualistic schemes (compensation schemes where pay depends only on the individual's own performance) and thus must be at least as good as such schemes, determining circumstances under which simple relative performance schemes (such as contests) do better than the simple kinds of individualistic schemes (such as piece rates) often found in practice is a far more difficult question. We show that contests will be preferred to (even nonlinear) individualistic schemes when the risk associated with the common environmental variable is large (relative to that of the idiosyncratic random variable). This is a theme to which we shall return later.

## II. Markets and Competition

Markets provide reward structures which have some of the properties of contests and relative performance schemes. The exact nature of the compensation scheme provided by the market for the owner-manager of a firm depends critically on the nature of the production technology and the market equilibrium. Consider, for instance, two firms engaged in cost reducing R&D. If the production technology is constant returns to scale, with marginal cost  $c$ , the profits of the  $i$ th firm will depend on his costs and those of his rival:  $\pi_i(c_1, c_2)$ . The profit function will differ markedly depending on whether the duopoly equilibrium is best described as a Nash-Cournot quantity setting equilibrium or as an Edgeworth-Bertrand price setting equilibrium. This is an example of a relative performance compensation scheme. But while our earlier study considered the *design* of a compensation scheme, here we are concerned with *describing* the consequences of the compensation scheme which is always *implicit* in the market equilibrium.

### A. Market Reward Structures Have the Property of Built-In Flexibility

If the cost functions facing different firms are correlated (there is a common environ-

mental variable affecting all firms) when costs are low, price is low, and therefore profits will be less variable and rewards more commensurate with the difficulty of the task than in a noncompetitive environment. To see this, consider the following modification of the previous example: let the costs of production be

$$(9) \quad c_i = k - \theta \mu_i,$$

where  $\mu_i$  is the (unobservable) level of managerial effort. There are a large number of firms, sufficiently large that they act as price takers. The "net" profit of the owner-manager  $\hat{\pi}_i$  taking into account the utility cost of managerial effort, is<sup>3</sup>

$$(10) \quad \hat{\pi}_i = [P - k + \mu_i \theta - V(\mu_i)] Q_i \\ = \pi_i - V(\mu_i) Q_i,$$

where  $Q_i$  is the output of the  $i$ th firm and  $P$  is the price of output. In competitive equilibrium

$$(11) \quad P = \underset{(\mu_i)}{\text{Min}} \{k - \mu_i \theta + V(\mu_i)\};$$

and

$$(12) \quad \hat{\pi}_i = 0, \text{ for all producing firms.}$$

Thus, competition *forces* each of the competitors to expend the correct amount of effort at cost-reduction activities; and it does this in such a way as to eliminate all variability in net profits (they are identically zero).

An owner-manager monopolist would have the correct incentives for cost minimization, but would face considerable variability in his profits.

### B. The Consequences of the Separation of Ownership and Management

But markedly different results obtain in the comparative performance of managers

<sup>3</sup>In this formulation, the effort expended is proportional to the level of output; there are neither increasing nor decreasing returns to scale in managerial technology. This assumption is made to avoid the difficulties which arise in comparing economies in which the number of managers (as opposed to aggregate managerial effort) is relevant.

who do not own all the resources which they manage. Assume that there is a competitive supply of managers. The equilibrium in the competitive market remains as described above. Since there is no risk, the optimal contract entails the manager receiving 100 percent of the profits at the margin. (Effectively, the identity of ownership and control is an endogenous characteristic of this economy.) But consider the problem of the monopolist attempting to hire a manager for his enterprise. He wishes to choose a contract which maximizes his own expected utility, subject to the constraint of being able to hire the manager. For simplicity, we assume the manager is risk averse, but the owner is risk neutral.<sup>4</sup> Then the optimal contract will entail some risk sharing by the original owner. (The interests of owner and manager do not coincide; the separation of the two is again an endogenous feature of this economy.) If we limit ourselves to simple linear contracts, then the pay of the manager will be of the form  $\alpha \pi_i + \beta$ , assuming that profits of the firm are observable, but the input of the manager is not. ( $\beta$  is a fixed fee, which may be either negative or positive, but plays no essential role in the subsequent analysis.) Three consequences follow immediately from the fact that it is optimal for  $0 < \alpha < 1$ .

1) Managers will not expend the efficient amount of resources on cost reduction; they will set  $\alpha \theta = V'(\mu)$ .

2) Managers will not adjust their effort to changes in circumstances as much as they would in the competitive regime; when it is easier to reduce costs, they effectively enjoy some of the benefits of the increased ease in greater leisure (to a greater extent than this would be true in a competitive economy):  $d\mu/d\theta = \alpha/V'' < 1/V''$ . This phenomenon is sometimes referred to as managerial slack. The argument for why noncompetitive environments may experience managerial slack is perhaps even stronger than we have put it here. The owner of the firm may have knowledge about the "normal" state of nature; the contract thus may implicitly specify a normal level of effort, and a normal expected return. When the level of effort required to attain

<sup>4</sup>All that is really required is that the manager be more risk averse than the original owner.

this "normal level" is greater, the managers have an incentive to present evidence to that effect; while if the level of effort required is less than this normal level, the managers have no incentive to present that information. The natural asymmetries of information give rise to an asymmetry in response to unusually good and unusually bad states.

3) Managers still have to bear some risk.

### C. Imperfect Competition

In the case of competitive economies, all the relevant information is embodied in the price; the owner of the firm does not have to base his manager's pay on the observed costs of his firm, say, relative to that of other firms. In the case of imperfectly competitive economies (for example, duopolies), the owner may wish to employ an incentive scheme which makes use of some of this detailed information, if it is available. Indeed, a slight modification of the first example given in this paper shows that it is possible with only two firms to design managerial incentive schemes in which the manager bears no risk yet has perfect incentives.

### D. Correlations Between Firms' Costs

The success of the relative performance schemes analyzed in the previous section was based on the fact that all firms faced identical cost functions. Assume that there are two firms, each of which has a choice of two technologies (or any linear combination of the two). If a firm devotes  $\lambda$  of its resource to technology  $i$ , its cost function per unit output will be  $F(\lambda, \theta_1, \theta_2)$  (where  $\theta_i$  are random variables). Define  $\lambda^*$  as the mixture which minimizes the expected costs,  $EF_\lambda(\lambda^*, \theta_1, \theta_2) = 0$ . Assume that the manager's pay depends on the difference between his costs and that of his rival. Clearly, if one firm imitates his rival, then costs are identical, and there is no risk. But if  $\lambda \neq \lambda^*$ , the manager can move  $\lambda$  towards  $\lambda^*$ , and decrease mean costs, and thus increase his pay. Though this increases his risk, for small deviations he acts in a risk neutral manner; thus *the Nash equilibrium entails an efficient choice of techniques and no risk*.

There are three important qualifications to this result. First, assume firms can only choose technique 1 or technique 2. The mean return with technique 2 is higher. If firm  $A$  chooses technique 1, the  $B$  manager's risk by choosing technique 2 may be so much higher that he isn't sufficiently compensated by the increased mean. Thus, there may be an equilibrium in which both firms choose technique 1, and an equilibrium in which both firms choose technique 2. One of these may Pareto dominate the other.

Second, if one firm has a comparative advantage in technique 1, and the other firm in technique 2, then each firm will not choose the technique which minimized its expected costs, but rather will choose a technique which is somewhere between the cost-minimizing technique and the technique chosen by the other firm.

Third, if there is idiosyncratic risk, then even when the two firms imitate each other, risk is not eliminated; still, if the two firms face the *same* stochastic technology, the only equilibrium is that where costs are minimized.

### E. The Anarchy of the Market Place: Excessive Competition

While the stories we have told here have pictured competition as reducing the risks faced by businessmen, businessmen often complain that unbridled competition forces them to bear an excessive amount of risk; they have, accordingly, often called upon the government to help regulate (stabilize) the market place. Some of these pleas are simply blatant attempts to cartelize the market, and to reap the monopoly rents which result. On the other hand, when the idiosyncratic risk is large, the variability in profits may be large.

In such situations, in our earlier studies of contests with risk-averse agents, relative performance schemes did not work well, since they imposed an excess amount of risk on the contestants. Similarly, in our earlier example of the development of the bomber, if the two researchers face different cost functions, then basing the compensation of one researcher on his performance relative to that of another imposes an additional source of

risk. The market imposes similar risks on managers, even when the compensation scheme is not based directly on relative performance, but on profits; because profits will, to a large extent, reflect differences in the costs of the firm relative to its competitors (or differences in relative performance in some other dimension, such as quality or marketing). To attempt to alleviate these risks by making pay less dependent on performance is likely simultaneously to ameliorate incentives.

### III. Conclusions

This paper has attempted to delineate a central role that competition plays: it allows the development of compensation schemes where pay is based on relative performance. Such compensation schemes have risk sharing, incentive, and built-in flexibility properties which make them superior to the best (individualistic) schemes which can be designed which do not make use of such information. The reward structures provided in competitive markets are, implicitly, related to relative performance. This provides an additional reason that competitive economies perform better than monopolies, a reason which is quite distinct from the loss in consumer's surplus arising from the monopolist's reduction in output. In particular, we have formulated a model in which monopolies are less efficient and less adaptable, and there is

more managerial slack than in competitive economies. (In spite of the widespread belief that monopolies are less efficient than competitive firms—including the intertemporal inefficiencies arising from inadequate allocation of resources to *R&D*—in traditional neoclassical models, there is no managerial slack and monopolies are perfectly efficient.) We have indicated that there are limits to the extent to which the market may reduce risks: in some cases, competition may effectively increase it. An examination of the full consequences of our observations, including an investigation of the constrained optimality of the economy and the implications of our analysis for policy, must await another occasion. What should be apparent is that the perspectives into the functions of competition in market economies arising from the approach taken here stand in market contrast to those provided by the traditional competitive paradigm.

### REFERENCES

- Lazear, Edward P. and Rosen, Sherwin, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, October 1981, 89, 841-64.
- Nalebuff, Barry J. and Stiglitz, Joseph E., "Prizes and Incentives: Towards a General Theory of Compensation and Competition," *Bell Journal of Economics*, forthcoming.

## THE BUDGET AND INFLATION

### Budget Expansion and Cost Inflation

By ASSAR LINDBECK\*

It has been increasingly recognized in recent years that expanding government budgets may contribute to "cost inflation," that is, inflation that does not require high or rising levels of capacity utilization. A basic idea is that higher taxes, to finance increased government spending, are directly or indirectly shifted onto product prices—a phenomenon that has been baptized "tax-shift inflation."

Not only are various types of indirect taxes, such as sales taxes, value-added taxes, payroll taxes, taxes on intermediary inputs, etc., assumed to be shifted onto the prices of final output, it is also assumed that private agents try to compensate themselves, in the form of *higher nominal factor prices*, for reductions in real after-tax income due to actual or expected tax increases. Indeed, in the case of higher (expected) indirect taxes on consumer goods, this idea is already embedded in the expectations-augmented Phillips curve. Though tax-shift inflation may occur also in the case of increased income taxes, there is an important difference between the two cases. With higher indirect taxes (and possibly also taxes on profits, as constructed in reality), firms first experience (or expect) higher production or sales costs, which are shifted onto product prices. These in turn create demands for higher factor prices, wages in particular, in a *second round*. With higher income taxes for employees, by contrast, shifting onto higher product prices presupposes tax shifting onto higher factor prices (wage rates), at the outset, with second-round effects on product prices.

Regardless of whether a tax shift on factor prices, induced by an expansion of the government budget, comes from the reactions of individual employees in some submarkets (for labor), or from more aggressive bargaining by unions in other submarkets, the phenomenon may be depicted as an upward shift of an aggregate supply curve for output in price-output space. Thus, a tax-shift impulse as the government finances its increased spending strengthens the depressive effect on output, and moderates the deflationary effect on prices.

The likelihood that such events will generate a process of continuing cost inflation, rather than a once-and-for-all increase in the price level, is, of course, greater for recurrent expansion of the government budget than for a once-and-for-all increase in the budget. Moreover, for such a process to lead to substantial price increases without heavily reduced output, not only monetary accommodation and/or higher labor demand by the government sector, but also a depreciation of the currency, is necessary (if other countries do not inflate at about the same rate).

To probe deeper into the relation between budget expansion and cost inflation, it is useful to take a closer look at the behavior of individual employees, unions, voters, and the government, and the interrelation between them. Indeed, that is the main purpose of this paper.

#### I. Reactions of Individual Employees

When looking at individual employees, a simple but useful microeconomic framework is to assume that the individual maximizes a well-behaved utility function with one private good, one publicly provided good (public consumption), and leisure as arguments. In principle, increased income or consumption

\*Institute for International Economic Studies, University of Stockholm. I am grateful for comments on an earlier version from, in particular, Lars Calmfors, Thorvaldur Gylfason, and Torsten Persson.

tax rates, combined with higher public spending (on goods and services), would then create a negative substitution effect on the supply of labor (in the official economy), while the positive income effect on labor supply of the tax increase itself would be counteracted by a negative income effect of the higher public spending.

However, empirical studies do not suggest very large substitution effects for males. And for females, where the substitution effect of changes in tax rates seems to be substantial, publicly supplied or subsidized goods—education, health care, old age care, and child care, etc.—are often close substitutes for leisure, or rather traditional home work of females. This means that the “direct” substitution effects against labor supply of women of the higher tax rates are often counteracted by positive cross-substitution effects on the labor supply of this group by way of the increased public provision (or subsidization) of goods and services (see my 1982 article).

The conclusion is that it is difficult to build a *strong* case for tax-shift inflation on the adjustment by individual employees of the number of hours of work supplied, and therewith connected upward shifts of the aggregate supply curve of labor, and hence indirectly also the aggregate supply curve of output.

However, there is an additional route for cost-inflation effects via the behavior of individual employees: negative effects on productive effort. The most obvious possibilities are that higher marginal tax rates create disincentives for the intensity of work, the ambition to strive for promotion, the willingness to shift from one job or geographical region to another, the desire to invest in human capital (if taxes are progressive), etc. Both households and firms may also be increasingly induced to substitute productive effort in the “underground economy” for work in official markets—as well as to participate increasingly in zero sum games of search for tax loopholes and profitable financial speculation. The ensuing slowdown of productivity growth in official markets is likely to result in higher inflation in a short-

and medium-term perspective (for instance less than five or ten years), as wage demands are probably partly based on previous experiences of productivity growth.

## II. Reactions of Labor Unions

Further light is shed on tax-shift inflation if the focus of the analysis is moved from individuals in equilibrating labor markets to wage demands of labor unions in collective bargaining. A useful microeconomic framework for the analysis of this issue is that each labor union tries to maximize a preference function with the real after-tax wage rate, public consumption, and the level of employment as arguments.

Let us then for the time being keep public consumption *exogenous*, and assume that the union chooses an optimum position between the real after-tax wage rate and employment, the latter being a negative function of the before-tax real wage rate. (We may then assume that the wage rate is set at a level above the market-clearing rate.) A rise in the tax rate, combined with higher public consumption, will then (using cardinal language) increase the marginal utility of the after-tax real wage rate relative to employment and public consumption. We would then predict an attempt by the union to restore the after-tax real wage rate, more or less, by trying to push up the nominal (and real) before-tax wage rate at the expense of employment. This effect is strengthened if private consumption (or rather after-tax real wage rates) is a (gross) complement to public consumption in the union utility function, but weakened if it is a (gross) substitute. The conclusion is that, except for cases of strong substitutability between private consumption and publicly provided goods (in the union utility function), a combined increase in taxes and publicly provided goods would be expected to induce attempts by unions to raise before-tax nominal and real wage rates. If wage rivalry among unions is pronounced, unions that have originally attempted only modest tax shifts on wages may ask for compensation for the higher tax shifts attained by others. If firms then respond to

higher wage rates by raising product prices, a combination of a wage-wage and a wage-price spiral easily emerges.

Obviously, not only *increased* tax rates, but also *high* tax rates, may contribute to cost inflation. In particular, in a society with high marginal tax rates, very large money wage increases are necessary to achieve even quite modest increases in real disposable income, or to compensate for even quite modest price increases. These tendencies are accentuated if the tax system is heavily progressive and not fully indexed. One rather well-known inference is that although high marginal tax rates, and a progressive tax system, create "automatic" built-in stabilizers in the economy on the demand side, they also tend to create built-in destabilizers on the cost side. Another inference is that for a country with a highly progressive tax system, where organized groups are involved in after-tax real income bargaining, it is very dangerous, from the point of view of inflation, to "let in" foreign price increases via international trade. Such a country would have to revalue at early stages of international inflationary periods to prevent strong "cost inflation multiplier effects" on the domestic economy of foreign price increases.

### III. Union-Government Interaction

The further macroeconomic consequences of tax-shift inflation impulses depend in a crucial way on the reaction of the government—and the feedback of government behavior on the behavior of unions (see Thorvaldur Gylfason and myself, 1982). For instance, one would expect a Keynesian-type government to validate the inflationary tendencies in an attempt to guarantee full employment by way of accommodating fiscal, monetary, and exchange rate policies. In some countries, the government may react to the stagflationary tendencies simply by expanding employment in the public sector (see Hans Söderström and Staffan Viotti, 1979), thereby introducing an element of "reverse (but still positive) causation" between public sector expansion and inflation. Indeed, not only the rate of expansion, but

also the *size*, of the government employment sector may stimulate cost inflation. For if unions try to push up wages to the point where the employees risk their jobs to a considerable extent, wage bargaining would be expected to be more aggressive in the government sector than in the private sector, because of the stronger job guarantees.

When the unions learn that the government responds with monetary or fiscal expansion to unemployment tendencies in the private sector, the demand curve for labor, as perceived by unions, becomes steeper. As a result, unions would normally be stimulated to strengthen their attempts to push up wage rates (see Lars Calmfors and Henrik Horn, 1982). However, it is possible that the government, too, and not just the unions, adjusts its behavior patterns due to learning. For instance, if the electorate becomes "unhappy" enough with rapid, and possibly also rising, inflation, the government might sooner or later feel compelled to shift to a less accommodating budget policy, with higher unemployment as a result. Such a shift in policy is particularly likely if the government hopes that the unemployment level can be brought down again later, possibly before the next election (for instance, via lower union wage demands).

### IV. Voter-Government Interaction

The considerations above emphasize that to understand tax-shift inflation, it is crucial to begin with a realistic model of both economic and political decision making, and hence to elucidate the driving forces behind both the expansion of the government budget and actual macroeconomic policies.

The "traditional" explanation for higher public spending in a growing economy is, of course, increased demand for infrastructure and public provision of goods with strong externalities, collective goods being the extreme case. However, this explanation is certainly not enough to explain the rapid expansion of public budgets during the last one or two decades. Indeed, an additional factor has already been mentioned: increased public employment, or subsidies to private em-

ployment, to "mop up" employees who have been pushed out of the private sector (for instance, because of aggressive wage behavior of unions). However, a basic hypothesis in this paper is that the extraordinarily rapid recent expansion of the government budget in several countries, particularly perhaps in northwestern Europe, is largely a result of the increased use of government budget policies for attempted redistributions of income and welfare. This is indicated already by the *types* of spending that have increased the most, that is, transfer payments, Social Security payments, and government subsidies and/or the public provision of specific "noncollective" services such as education, medical care, child care, and old age care. (The gradual relative price increases of such services have strongly accentuated the expansion of the ratio of public services to *GDP* at current prices.)

These redistribution policies have influenced not only the *size distribution* of (yearly and lifetime) income, as predicted for instance by the median voter theorem of public choice. We have also experienced very active attempts by governments to reshuffle income within the *horizontal distribution* of income and welfare, that is, between various socioeconomic groups in society, such as families that differ with respect to the number, age, and productive capacity of the members (reflecting different numbers of children and pensioners); citizens with different professions, working in different industries or living in different regions; people living in different types of housing (owner-occupied or apartment houses); people holding different types and amounts of assets and debt; as well as people with different preferences with respect to various consumer goods and leisure activities, etc.

A fundamental explanation why the redistributions actually attempted have resulted in a rapid expansion of the government budget is that while taxes are often rather *general*, benefits from the government (transfers as well as public services) are often quite *specific*, that is, they are tied to special socioeconomic groups and age classes such as those mentioned above. This makes it possible for politicians to "buy votes" from

each special group at a time on the basis of the explicit or implicit assumption that the accompanying tax increases are paid mainly by others. It is then important to point out that this "prisoner's dilemma" behavior of voters does not imply any irrationality on their part—as long as coalitions of the various subgroups in society are difficult, or costly, to organize. A rapid expansion of the government budget is often also facilitated by the absence, in most countries, of constitutional rules that require qualified majority (à la Wicksell) or incite referenda about public spending programs, or that force parliament to tie a decision about the financing to each separate spending increase (as conceived in Musgrave's vision of a separate "Allocation Branch" in the Treasury Department). The same factors make it difficult to *cut* government spending, as that would hit specific groups, without creating large benefits for the average tax payer; "ratchet effects" then easily emerge.

A reason why attempted redistributions of this type have become more important in recent decades is that the rigid class society, with strong class loyalty in voting, has been replaced by a society with highly fragmented (organized as well as nonorganized) interest groups, implying that every political party can buy votes from "everybody." These activities of the political parties have been further stimulated by the drastic fall in the marketing costs of party platforms because of developments in the mass media, in particular, the increased importance of television.

A theory of this type not only predicts a rapid rate of expansion of the government budget, it also predicts that each group of citizens becomes disappointed and frustrated when they find out that the tax burden *in fact* goes up considerably for them as well—simply because other groups in society, too, are getting specific transfers and benefits from the government. One conceivable reaction of citizens is to cease supporting the politicians who play this game, and instead vote for politicians who promise to stop the "leapfrog" expansion of public spending. However, another and perhaps more likely reaction is that voters become even more

anxious to accept offers from politicians of specific benefits, because they want to keep up with the benefits received by others. Moreover, the ever-higher marginal tax rates make benefits from the government, which are often tax free, much more attractive than income earned from additional work which does not yield much higher disposable income, at least not in the most advanced welfare states of Western Europe. Increased political lobbying for benefits by organized interest groups would then also be expected to increase.

The predicted macroeconomic result is then further rounds of budget increases, with tax-shift inflation via the behavior of both individual employees and labor unions, as described above. As a consequence, the earlier mentioned wage-wage and wage-price *spirals* of cost inflation, in connection to expanding government budgets, may be widened to include also strong elements of voter-government interaction, leading to further expansion of the government budget. The macroeconomic development in the Scandinavian countries during the 1960's and 1970's may serve as an illustration.

However, let us take one additional step in the analysis. In a long-run perspective, a rapid expansion of the government budget may have even more fundamental macroeconomic effects than spirals of competing income claims and cost inflation. In a later stage of the process, it is not unlikely that large *budget deficits* will emerge, either because of public resistance to increased tax rates or because the tax base of privately generated income is reduced by high product wage rates that keep down employment and output, disincentive effects on productive work, and/or restrictive monetary policy by the government to reduce inflation. The higher burden of the public debt as well as increased budget costs for the unemployed and for increased employment in the public sector, will accentuate the deficits. Moreover, if the product-wage rate becomes high relative to labor productivity, the current account of the balance of payments easily turn into deficit too (see Söderström-Viotti).

Thus, a situation may emerge with a combination of a rapid expansion of the govern-

ment budget, stagflation, slow productivity growth, and large deficits in both the current account and the government budget, where the deficits may be regarded as "structural," rather than "cyclical," in the sense that they are generated by a combination of a trend-wise retardation of output growth and continued political pressures for a rapid expansion of government spending. If the budget deficit is financed mainly by borrowing in domestic capital markets, the ensuing interest rate increase would be expected to reduce private investment, with further negative effects on productivity growth. Indeed, after the mid-1970's a development of that type seems to have taken place in a great number of countries, though the reasons are probably much more complex than those mentioned here; for instance, other types of supply shocks than higher product wage rates, and the accentuation of inflationary inertia as well as restrictive government policies induced by these developments, are obvious examples of additional explanations.

## V. The Empirical Relevance

Even though I have given several reasons to expect a positive and causal connection between the size and rate of increase of the government budget on the one hand, and cost inflation on the other, that, of course, does not mean that we should expect a *simple* statistical correlation between these phenomena. Indeed, cross-country regressions from the post-World War II period for highly developed countries do not reveal any positive association between inflation and the size, or rate of expansion, of the government budget. The conclusion is that country differences in the rate of inflation, at least so far, have been dominated by other dissimilarities between countries that are relevant for inflation, such as domestic capacity utilization, the strength of the full-employment guarantees, monetary and exchange rate policy, the behavior of labor unions, etc. Moreover, in a world with rather fixed exchange rates, as before the early 1970's, the long-term inflation trend in individual countries is to a considerable extent tied to the inflation trends of the world economy,

rather than to domestic demand and cost-push factors.

Nevertheless, studies of the interaction between the size and rate of expansion of the government budget on the one hand, and cost inflation on the other, may increase our understanding of the *mechanisms* of inflation in individual countries even though other influences probably have been more important for the *rate* of inflation—at least up till now.

#### REFERENCES

- Calmfors, Lars and Horn, Henrik, "Employment Policies and Centralized Wage Setting," Seminar Paper No. 238, Stockholm: Institute for International Economic Studies, 1982.
- Gylfason, Thorvaldur and Lindbeck, Assar, "The Macroeconomic Consequences of Endogenous Governments and Labor Unions," Seminar Paper No. 232, Stockholm: Institute for International Economic Studies, December 1982.
- Lindbeck, Assar, "Tax Effects Versus Budget Effects on Labor Supply," *Economic Inquiry*, October 1982, 20, 473-39.
- Söderström, Hans Tson and Viotti, Staffan, "Money Wage Disturbances and the Endogeneity of the Public Sector in an Open Economy," in A. Lindbeck, ed., *Inflation and Employment in Open Economies*, Amsterdam: North-Holland, 1979.

# The Interconnection Between Public Expenditure and Inflation in Britain

By WALTER ELTIS\*

Two lines of argument link public expenditure and inflation. The first focuses attention on that part of public expenditure which is not financed by taxation, and attributes inflation to the financing of consequent deficits by methods which involve increases in the money supply. The second, in contrast, focuses on the fraction of public expenditure which is financed through taxation and partly attributes inflation (or the higher unemployment needed to contain it) to the extra wage increases which workers seek in their efforts to obtain the rate of growth of net of tax real incomes which they have come to expect. The relevance of these two theoretical approaches to the acceleration of Britain's inflation rate will be considered in turn.

Table 1 shows why there is a strong a priori possibility that widening budget deficits may have contributed to the acceleration of inflation, and the broad statistical correspondences indicate how readily econometric work has been able to establish results which link the acceleration of inflation to growth in the money supply, and which attach considerable weight to the widening of the budget deficit in the explanation of the latter. There are however difficulties with a conventional interpretation of the data along these lines.

The £M3 series increased 20 percent per annum from 1971 to 1973 when government borrowing was just 2.3 and 3.3 percent of GDP, while when the need for borrowing peaked at 6.8 and 9.5 percent of GDP in 1974 and 1975, £M3 increased by just 8.6 and 6.8 percent. The £M3 explosion therefore began well before the budget deficit became difficult to finance, and growth in £M3 actually became quite slow while the deficit was at its peak. There is the further difficulty that neither £M3 nor M1 offers

consistent support for a conventional interpretation of the relationship between the money supply and the subsequent behavior of the money GDP. Thus £M3 provides the principal monetary explanation of the inflationary explosion of 1973-74 because M1 increased quite slowly in 1971 and 1972, but it was M1 and not £M3 that increased slowly in Mrs. Thatcher's first two years when unemployment rose sharply, sterling soared, and all the symptoms of monetary tightness were present despite a rate of growth in £M3 of 16.0 percent. Attention is therefore now moving to PSL2, a wider measure of liquid assets, and to M2 in the hope that consistent and stable monetary relationships may emerge.

A further difficulty with a simplistic linkage between the budget, the money supply, and the acceleration of British inflation is that as soon as the real capital gains and losses which result from inflation are allowed for, Bank of England economists estimate that British governments which had apparent borrowing requirements of 8.6 percent of GDP in 1973-75, 6.2 percent in 1976-78, and 7.5 percent in 1979, were actually repaying debt in real terms in those years (C. T. Taylor and A. R. Threadgold, 1979). That British governments were not adding to real debt while apparent deficits were so large throws some doubt on the proposition that the rapid monetary growth which occurred was due to a reluctance or inability to market bonds.

A final difficulty with the straightforward proposition that widening budget deficits led to accelerating monetary growth and a consequent trend acceleration of inflation is that prices and the exchange rate may often depend on domestic wage increases, the money supply then being increased to adjust to these. In a world of this nature, wage inflation which will be influenced by disappointment

\*Fellow of Exeter College, Oxford University.

TABLE 1—BRITISH GOVERNMENT BORROWING AND INFLATION

	Percentages						
	1957	1961	1965	1969	1973	1979	1981
<i>GDP</i>							
Government Expenditure	34.4	34.6	36.2	39.1	40.0	42.3	45.3
Taxation	29.3	29.5	30.8	36.8	33.0	35.0	39.0
Government Borrowing	2.4	2.7	3.4	−0.6	5.0	6.3	4.5
	Annual Percentage Increases						
	1957–61	61–65	65–69	69–73	73–79	79–81	
<i>M1</i>			2.8	11.6	12.8		8.2
<i>£M3</i>			5.8	15.4	11.9		16.0
<i>PSL2</i>			7.2	14.2	12.3		13.0
<i>GDP</i> at Market Prices	5.5	6.9	7.0	11.9	17.6		13.3
Market Price <i>GDP</i> Deflator	2.3	3.6	4.4	8.0	16.1		15.7

Source: Derived from various tables in *National Income and Expenditure* and *Economic Trends*, HMSO, London. Government Expenditure is total expenditure less government interest and dividend receipts. Government Borrowing does not equal expenditure minus taxation because general government has other sources of income such as rents.

with the rate of growth of net of tax real incomes, among several considerations, will lead to price acceleration. The time has therefore come to discuss the second link between public expenditure and inflation.

It has been widely argued in Britain that wage equations should make some allowance for workers' aspirations to increase or maintain their real *net of tax* earnings. Perhaps the simplest form for such an equation is

$$(1) \quad \dot{w} = {}_e\hat{p} + {}_e\hat{a} + \partial T_d / (1 - T_d) + f_1(u_n - u), \quad f'_1 > 0;$$

namely that  $\dot{w}$ , the "annual" rate of increase in money wages equals  ${}_e\hat{p}$ , the expected rate of inflation, plus  ${}_e\hat{a}$ , the expected annual increase in net of tax real incomes, plus  $\partial T_d / (1 - T_d)$ , compensation for increases in  $T_d$ , the supposedly uniform rate of direct taxation, plus a term,  $f_1(u_n - u)$ , which recognizes that money wages also depend on the difference between actual unemployment and  $u_n$ , unemployment in a steady state or the natural rate. Where  $u$  exceeds  $u_n$ , wage increases are depressed, both because workers acquiesce in smaller increases, and because employers are unable or reluctant to grant substantial increases where demand is weak.

There is also a price equation of the form:

$$(2) \quad \hat{p} = \hat{w} - \hat{q} + (\widehat{1 + T_i}) + f_2(u_n - u), \quad f'_2 > 0.$$

This equation has been simplified to exclude the influence of import prices on domestic prices, and it is assumed that  $\hat{p}$ , the rate of price inflation, equals  $\hat{w}$  less  $\hat{q}$ , the annual rate of increase in productivity, plus  $(\widehat{1 + T_i})$  to show the effect on price inflation of increases in the uniform rate of indirect taxation, plus  $f_2(u_n - u)$  to indicate that cost increases are less readily passed on where unemployment exceeds the natural rate.

When (1) is used to substitute for  $\hat{w}$  in (2),

$$(3) \quad \hat{p} = {}_e\hat{p} + {}_e\hat{a} - \hat{q} + (\widehat{1 + T_i}) + \partial T_d / (1 - T_d) + f_3(u_n - u),$$

where

$$f_3(u_n - u) = f_1(u_n - u) + f_2(u_n - u).$$

Superficially, (3) is a Keynesian equation for price inflation which depends on tax increases and the inflation rate that workers expect to be compensated for, and the real income increases they hope to obtain, tempered merely by some dampening effect if unemployment exceeds the natural rate. Sup-

pose that this "Keynesian" equation is supplemented by the assumption of a most rigid connection between the increase in effective demand in money terms,  $\hat{y}$  (which equals  $\hat{p} + \hat{g}$ , the rate of increase in real output), and  $\hat{M}_{-1}$ , the rate of increase in the money supply in the previous "year" (the uniform lag assumed for simplicity), so that

$$(4) \quad \hat{g} = \hat{M}_{-1} - \hat{p}.$$

This very simple and rigid money demand equation which the data of Table 1 suggests is by no means contradicted by British developments (it is the relevance of budget deficits to  $\hat{M}$  and the appropriate measurement of this which creates difficulties), is perfectly compatible with the superficially Keynesian (3). If this equation generates a  $\hat{p}$  which exceeds  $\hat{M}$ , then (4) indicates that  $\hat{g}$  will be negative, the economy's real output will fall drastically, unemployment will soar in relation to the natural rate, and this will make the  $f_3(u_n - u)$  term in (3) increasingly negative until the rate of inflation falls into line with the rate of growth of the money supply. The appropriate long-term inflation rate derived from this will be,  $\hat{M} - \hat{q}$ , if the labor force is constant so that  $\hat{g} = \hat{q}$ , where  $u = u_n$ .

The need for unemployment to exceed the natural rate to bring any inflation generated through (3) into line with the rate compatible with monetary expansion as set out in (4) would disappear if expectations were truly rational. In that event,  ${}_e\hat{p}$ , the expected rate of inflation, would equal  $\hat{M}_{-1} - \hat{q}$ , and  ${}_e\hat{a}$ , the expected rate of increase in net of tax real incomes, would equal  $\hat{q} - (\widehat{1 + T_i}) - \partial T_d / (\widehat{1 - T_d})$ , so workers would bargain for higher wages in the knowledge that they could not hope for long-term compensation for price rises which were out of line with the growth of the money supply; and they could not expect to raise their net of tax real incomes faster than the rate of productivity growth less the extra resources which go to the public sector through higher taxation. With these substitutions in (3):

$$(5) \quad \hat{p} = \hat{M}_{-1} - \hat{q} + f_3(u_n - u);$$

and from (4) and (5),

$$(6) \quad \hat{g} = \hat{q} - f_3(u_n - u).$$

Thus with the assumption of rational expectations, output will actually rise faster than productivity where unemployment exceeds the natural rate, and more slowly where it is less than this, so there is a stable convergence of unemployment to the natural rate.

Sadly, in Britain at any rate, there has been no basis for a rational expectation that prices will rise in line with government announced monetary targets, less productivity growth. Even if there were a rationally inclined trade union movement, this could not have the perception to predict when prices would move in line with  $\text{£}M3$  or  $M1$ , not to mention  $M2$  and  $PSL2$ , and in the absence of a measure of the money supply which predicts future movements in the money National Product satisfactorily and consistently, there is no theoretical reason why inflationary expectations should keep in line with monetary targets. Adaptive expectations may therefore be the best practical assumption for the expectations which influence wage behavior, and the simplest form for these is  ${}_e\hat{p} = \hat{p}_{-1}$ , that is, wage bargainers expect prices to continue to rise at the rate at which they have been rising in the immediate past. It can also be assumed that they will expect net of tax real incomes to continue to grow at the rate at which they have been rising over a past time horizon of  $H$  years, so that  ${}_e\hat{a} = \hat{a}_{-1 \dots H}$ . With these substitutions, (3) becomes

$$(7) \quad \hat{p} - \hat{p}_{-1} = \hat{a}_{-1 \dots H} - [\hat{q} - (\widehat{1 + T_i}) - \partial T_d / (1 - T_d)] + f_3(u_n - u),$$

and from (1), and (2) with the adaptive expectations substitutions,

$$(8) \quad \hat{w} - \hat{w}_{-1} = \hat{a}_{-1 \dots H} - [\hat{q}_{-1} - (\widehat{1 + T_i}) - \partial T_d / (1 - T_d)] + f_3(u_n - u).$$

Equations (7) and (8) indicate that, with

adaptive expectations, there are three circumstances where price and wage inflation will accelerate. As is well known, it will accelerate if  $u$  is less than  $u_n$ . It will however also accelerate if  $\hat{a}_{-1 \dots H}$ , the rate of increase in net of tax real incomes over the previous  $H$  year time horizon exceeds the rate of productivity growth less the rate at which taxes are rising. Thus either a slowdown of productivity growth, or a faster increase in taxation than workers have become familiar with in the previous  $H$  years, will produce an initial acceleration of wage and price inflation as wage bargaining continues to be based on the previous rate of real income growth. A faster rise in import prices than in domestic prices (which is not specifically allowed for in the equations) will clearly also produce accelerating inflation through disappointed expectations.

This accelerating inflation as a result of disappointed expectations may then come into conflict with (4), the monetary demand equation. If governments simply allow the money supply to track wages upwards, and this was the British practice until the mid-1970's, the money supply will accelerate in line with (7) and (8), and there will then be no difficulty in reconciling the adaptive expectations wage and price equations with the availability of money. If, however, the rate of growth of the money supply in (4) depends on inflationary targets, and this has been the situation in Britain since 1976, then an unexpected falling off in productivity growth or unexpectedly rapid increases in taxation will set off an acceleration of inflation which will then, through (4), leave insufficient monetary demand to allow output to rise as fast as productivity. Unemployment will then rise in relation to the natural rate until inflation falls into line with the monetary targets, or, alternatively, governments will succumb to pressures to relax their monetary targets in order to accommodate the acceleration of inflation which is indicated by (7) and (8).

The inflation rate will rapidly come into line with the monetary targets if either of two possible conditions holds. First, if the time horizon over which workers' expectations about real income increases are formed is brief, wage and price acceleration as a

result of tax-push or productivity disappointment will be rapidly damped out. The rate of increase in money wages will then adjust downwards rapidly to new situations where real net of tax incomes can no longer grow. Tax-push inflation, etc., will also be rapidly damped out if the  $f_3(u_n - u)$  term in (7) and (8) is strong so that wage and price inflation are very sensitive to a small excess in unemployment over the natural rate. Sadly, it seems more likely that, in Britain, the period over which expectations about real income growth are formed may be quite long, and wage and price increases which evidently show some sensitivity to divergences between actual unemployment and the natural rate may be less sensitive to this than in some other economies where rigidities due to unionisation are less pervasive. This means that British wage and price inflation may, as a result of disappointed expectations, significantly exceed the rates compatible with the rate of growth of the money supply, and only come into line with monetary targets at a cost of high and prolonged unemployment. This is arguably the mechanism through which Britain's rising tax ratio and disappointing productivity growth has added to the stagflation which has also been partly due to the unexpected commodity price explosion in the early 1970's, and to the adverse effects of rising marginal taxation on supply, which has an influence on  $\hat{q}$  in the equations.

Table 2 illustrates the applicability of the theory to Britain. The data show that from 1957 to 1965, real output available per worker increased by 2.2 percent per annum and real wages net of tax at 2.6 percent. It is therefore not implausible that by 1965,  $\hat{a}_{-1 \dots H}$ , the expected rate of growth of real net of tax incomes had become something like 2.5 percent. While productivity advanced still faster in 1965-69, taxation increased sharply with the result that in these four years, what was available net of taxation increased by only 0.8 to 1.2 percent per annum (depending on the method of estimation used). Table 1 showed that fiscal and monetary policies were extremely conservative in this period, but the wage rate of the average manual worker (which had increased at an average annual

TABLE 2—THE GROWTH OF REAL PER CAPITA INCOMES NET OF TAXATION

	Annual Percentage Rates of Growth			
	1957-65	1965-69	1969-73	1973-79
Real <i>GDP</i> per Worker	2.2	3.1	3.6	1.1
Plus Gain from Terms of Trade	0.3	0.0	-0.3	0.0
Less Diversion to Higher Taxation	-0.3	-2.3	1.5	-0.5
Real Output Available per Worker	2.2	0.8	4.8	0.6
Increase in Real Wage Net of Taxation <sup>a</sup>	2.6	1.2	3.7	0.9

Source: The first three lines are derived from *National Income and Expenditure and Economic Trends*, while the final line is derived from Bacon's and my 1976 book, Table 6.1, extended to cover 1957 and 1979.

<sup>a</sup>Shown for comparison.

rate of 3.7 percent from 1957 to 1965) increased by 4.6 percent in 1965-66, 3.9 percent in 1966-67, 6.6 percent in 1967-68, 5.3 percent in 1968-69, and by 10.8 percent in 1969-70. Wage increases therefore accelerated from an annual rate of around 4 to around 10 percent, despite the complete elimination of the budget deficit by 1969 and the very modest rates of monetary growth shown in Table 1. This acceleration of wages as a result of a fall in the rate of growth of real net of tax per capita incomes from about 2.5 to about 1 percent is precisely what equation (8) would predict. The 1967 devaluation of sterling is an alternative explanation of this acceleration of wage inflation, but retail price inflation only reached 5.4 percent by 1968-69.

In the subsequent cycle of 1969-73, supply-side tax cuts (as they would now be called) amounting to 3.8 percent of *GDP* offered the prospect of a dampening of inflation with real net of tax incomes growing at unprecedented rates if there had not also been an excessive pressure of demand in relation to supply. In this cycle, the rate of growth of the money supply which was only partly due to the widening budget deficit was far faster than in the previous cycle, which plausibly raised output rapidly through a monetary demand equation such as (4), since the rate of increase in £M3 (which has been universally regarded as the most significant measure of the money supply in this cycle) was far faster than the initial rate of inflation. The rapid growth of output which fol-

lowed from the initial excess of monetary growth over the damped rise in prices reduced unemployment to well below the natural rate by 1973. In the 1950's, falling rates of inflation had been achieved with unemployment rates of around 2 percent, but unemployment benefits were raised from 35-40 percent of after tax average earnings in the 1950's to around 75 percent of average earnings in the early 1970's, and it is arguable that the natural rate of unemployment rose sharply in the 1960's for this and other reasons, including the increase in trade union militancy and power and the extra wage rigidities this introduced which was associated with the very slow growth in real net of tax earnings in the early 1960's (Robert Bacon and myself, 1976). The British unemployment rates of 2½ percent which were reached in 1973 and in 1974 were well below the now far higher natural rate, and the reduction in unemployment to an unsustainable level arguably contributed significantly to the acceleration of inflation in 1972-74. This was reinforced by the money-induced fall in the exchange rate of sterling at the start of the floating rate period. Table 2 suggests that tax-push considerations and the unexpected increase in the rate of productivity growth could have contributed to a deceleration of inflation in 1969-73 if the influence of excess demand on wages and prices had not been so pronounced. After 1972, the commodity price explosion added to the acceleration of inflation, but all countries suffered from this and the far faster increase

in inflation in Britain than in other Western economies owed much to the influence of the highly inflationary monetary and fiscal policies which were pursued.

Wage and price inflation rose to around 30 percent at the start of the cycle of 1973-79. Inflation was accelerating at the end of the previous cycle, and the sharp and entirely unexpected slowdown in productivity growth accentuated the force of the inflationary cost and price interactions in the manner set out in (7) and (8) above. The temporary dismantling of most wage and price controls in 1974 which allowed previous repressed inflation to unwind added to the acceleration of inflation. This was checked in 1975, but halting the acceleration of inflation required a rate of growth of the money supply which was so much less than the rate of increase in prices that the growth of real demand (from a monetary demand equation such as (4)) fell far below the rate of increase in productivity which sent unemployment soaring to well above the natural rate. From this point onwards, stagflation was deep rooted, and the authorities had to choose between the intolerable unemployment needed to insert a sufficient brake in the wage and price equations and the enormous rates of monetary expansion which would allow money wages to compensate for expected inflation and the increase in real net of tax incomes which the British economy could no longer deliver, but

which workers nevertheless continued to expect for several further years. After 1977, it was only in 1982 when unemployment reached 14 percent that the annual growth in earnings fell below the target rate of growth of the money supply so that the real demand for labor could start to rise again. Squeezing out the unreal expectations of the early 1970's therefore required at least five years of excess unemployment.

It is to be noted that the frustration of expectations of real income growth which follow from rising taxation was only one of four possible causes of wage and price acceleration. Unexpectedly, slow productivity growth, a sharp adverse movement in the terms of trade, and unemployment below the natural rate also set off wage and price acceleration. The rising British ratio of taxation was therefore only one of several factors which contributed to the acceleration of inflation in the 1960's and the 1970's.

#### REFERENCES

- Bacon, Robert and Eltis, Walter, *Britain's Economic Problem: Too Few Producers*, London: Macmillan; New York: St. Martin's Press, 1976.
- Taylor, C. T., and Threadgold, A. R., "Real National Saving and its Sectoral Composition," Discussion Paper No 5, Bank of England, October 1979.

# Money, Credit Constraints, and Economic Activity

By ALAN S. BLINDER AND JOSEPH E. STIGLITZ\*

When government expenditures exceed current tax revenues, the resulting deficit must be financed either by issuing bonds, which imply obligations to levy future taxes, or by creating high-powered money. The choice between money and bonds is often thought to be of great moment for both real and nominal variables; that is, monetary policy matters.

There is by now a wide empirical consensus that monetary policy has effects on real variables like output and employment, but there is far less agreement about why this is so. The purpose of this paper is to take issue with some currently fashionable views of why money has real effects, and to suggest a new theory, or rather resurrect an old one—the loanable funds theory—and give it new, improved micro foundations.

## I. Some New Irrelevance Theorems

In classical monetary theory, prices are fully flexible and the future tax liabilities implied by government bonds are fully discounted. In such a world, government spending has identical effects whether it is financed by bonds (thus creating a “deficit”) or by current taxation, and an open-market purchase of bonds is equivalent to a money rain. Consequently, a swap of future for current taxes has neither real nor nominal effects, and a swap of money for bonds affects only the price level.

But these irrelevance theorems rest on micro foundations that are not well specified. For example, classical monetary theory presumably applies to a frictionless world of certainty and lump sum taxes, and mostly ignores the dynamic effects on real rates of

return that arise when monetary policy changes the expected path of the price level.

If an explicitly dynamic, general equilibrium model in which people form (rational) expectations about the uncertain future is constructed, a number of irrelevance theorems about government financial policy can be established, provided that financial changes do not redistribute the tax burden (see Stiglitz, 1981). For example, let the government reduce current taxes, issue bonds, and sometime later raise taxes to retire the bonds. Not only will such a policy leave real consumption and investment by all individuals in all states of nature unchanged, but neither will it change any prices. The reason is Say's Law of Government Deficits: the increase in the supply of government debt gives rise to an identical increase in the demand.

Other irrelevance propositions can be established. For example, if the government changes the maturity structure of its debt, or exchanges indexed for nonindexed bonds, such changes will be irrelevant because of exactly offsetting changes in the demands for different government securities. Similarly, a change in the rate of inflation that is matched by a change in the nominal interest paid on government debt does not disturb equilibrium in any market.

Some of these irrelevance results are familiar. Others contrast sharply with the implications of traditional portfolio theory. For example, a standard argument holds that a change in the maturity structure of the government debt will require a change in the term structure of interest rates to equilibrate the demands and supplies of different types of bonds. But this argument ignores the tacit, and exactly offsetting, changes in liabilities implied by the structure of taxes across time and states of nature. Perhaps individuals also ignore the implied tax changes. But to use this as a major theoretical underpinning of the effectiveness of monetary policy is to ground the theory in irrationality, an

\*Princeton University. We gratefully acknowledge financial support from the National Science Foundation and helpful discussions with Benjamin Friedman, Bruce Greenwald, Laurence Kotlikoff, and Andrew Weiss. A longer version of this paper appears as an NBER working paper.

TABLE 1—*F* TESTS OF IRRELEVANCE PROPOSITIONS<sup>a</sup>

	Equation explaining:		
	Nominal GNP Growth	Inflation Rate	Real GNP Growth
Hypothesis: All $b_i = 0$	6.9 <sup>b</sup> /9.2 <sup>b</sup>	3.5 <sup>c</sup> /2.8	2.6/4.0 <sup>c</sup>
Hypothesis: All $c_i = 0$	10.5 <sup>b</sup> /14.6 <sup>b</sup>	3.5 <sup>c</sup> /4.2 <sup>c</sup>	2.6/2.4

<sup>a</sup>Results from regressions with 3 lags are reported before the slash; results from regressions with 2 lags are reported after the slash.

<sup>b</sup>denotes significant at 1 percent level.

<sup>c</sup>denotes significant at 5 percent level.

anathema to economists of the Modern School.

## II. The Irrelevance of Irrelevance Theorems

If these irrelevance theorems are correct, then neither swaps between current and future taxes (nonmonetized budget deficits) nor open-market operations (creation of high-powered money) should matter.

To put these notions to a crude test, standard "causality" tests were run by regressing three critical U.S. time-series on their own lagged values, lagged values of changes in bank reserves, and lagged values of changes in government debt. Specifically, the regressions took the form:

$$\Delta X/X = a(L)(\Delta X/X) + b(L)(\Delta R/R) + c(L)(\Delta D/D) + e,$$

where  $\Delta$  is the first-difference operator;  $a(L)$ ,  $b(L)$ , and  $c(L)$  are polynomials in the lag operator;  $R$  is bank reserves;  $D$  is the government debt; and  $X$  is alternatively nominal GNP ( $Y$ ), real GNP ( $y$ ), or the GNP deflator ( $P$ ).<sup>1</sup> Regressions were run with the maximum lag set alternatively at two or three years.

If open-market operations were irrelevant, then changes in reserves should not "cause" any of the left-hand variables, once we control for changes in debt; that is, all the  $b$ 's

should be zero. In the case of nominal GNP, this hypothesis is easily rejected with  $F$  values of 6.9 and 9.2 (see Table 1). But for real GNP and prices, the evidence is mixed. In each case, one regression rejects the irrelevance proposition while the other does not.

If pure swaps between current and future taxes were irrelevant, then changes in debt should not "cause" any of the left-hand variables, once we control for changes in reserves, that is, all the  $c$ 's should be zero. The regressions for nominal GNP overwhelmingly reject this hypothesis (with  $F$  values of 10.5 and 14.6). And the regressions for inflation also reject it, though less decisively. However, we cannot reject the hypothesis that nonmonetized deficits are irrelevant for real GNP growth.

On balance, the evidence calls the strong forms of the irrelevance theorems into question and suggests a need to examine the assumptions that underlie them. Full rationality has already been mentioned. Equally obvious is the assumption that all taxes are lump sum; no one ever claimed that swaps among distorting taxes would be neutral.

The theorems also assume that taxes are distributionally neutral. It is well known that changes in the distribution of income and wealth across individuals can have real effects. Analogously, redistributing the tax burden across generations can have real effects if individuals have no heirs or fail to incorporate fully their heirs' welfare into their own utility functions. While these effects are probably present, one wonders about their empirical importance. Is redistribution across generations really the driving force behind monetary policy?

<sup>1</sup>Time was measured in fiscal years, and the sample period covered 1952–81. The term  $R$  is adjusted bank reserves, as calculated by the Federal Reserve Bank of St. Louis, and  $\Delta D$  is the increase in government indebtedness to the public during the fiscal year. For more detailed results, see Blinder (1982).

The irrelevance theorems also ignore the difference between interest-bearing government debt and non-interest-bearing money, which is held for transactions purposes. Traditional monetary theory has focused on this difference. Surely paper money and checking balances have advantages in transactions over other potential media of exchange. But are these advantages sufficiently large to explain the effectiveness of monetary policy by arguing, for example, that a contrived scarcity of the medium of exchange will constrain economic activity? In Italy, when there was a shortage of small change, candy became a medium of exchange. And now, with computerized banking, it should be relatively easy for velocity to change quickly to compensate for any shortage of money. Recent innovations like CMA's suggest that the transactions costs of providing a medium of exchange paying a market rate of interest cannot be very large. We believe that only regulation and lack of full rationality prevented checking accounts from paying slightly less than market interest rates for so long.

Another assumption pertains to the informational content of monetary or debt policy: the irrelevance theorems assume that policy actions do not change peoples' beliefs about the different states of nature. But if the government has superior information (which it does not make public), and uses this information in formulating policy, then policy might have real effects because of the information it conveys to the private sector. In addition, if monetary policy has a random element, individuals will have trouble distinguishing between price movements that are the consequence of real shocks and those that are the consequence of monetary shocks. This, too, can give money the power to influence real variables.

But can these informational issues be empirically important? We are skeptical. In addition to the weekly money supply number, a firm can look at its inventories, sales data, the national unemployment rate, and many other facts and figures that help it distinguish between real and nominal shocks. Besides, at low and moderate rates of inflation, people always know the current price

level within a very small margin of error, and therefore can easily convert any absolute price into a relative one with great accuracy. It therefore seems implausible that the issues emphasized by the new classical macroeconomics can rationalize sizable effects of monetary policy on output.

A final, and very critical, assumption that underlies the irrelevance theorems is that capital markets are perfect. But people cannot borrow freely at the government's interest rate, and for a very good reason: they might default. The probability of default, and the informational imperfections that it implies, lie at the heart of our alternative theory of how monetary policy works.

### III. Imperfect Information and Credit Rationing

Imperfect information about the probability of default has several fundamental implications for the nature of capital markets. First, it gives rise to institutions—like banks—that specialize in acquiring information about default risk. Such information is valuable. A lender with superior information can more easily distinguish between good and bad risks, thereby raising his own net (of default losses) rate of return. But such information is very specific (knowing that Company *A* is a good risk may tell us little about Company *B*) and, for a variety of reasons, is also difficult to sell.

Second, banks will devise nonprice mechanisms for screening out untrustworthy borrowers. As Stiglitz and Andrew Weiss have argued, reacting to excess demand for loans by raising the rate of interest may lower the bank's expected return because of adverse effects on the mix of applicants, and by inducing borrowers to undertake riskier projects. Thus credit rationing arises as an equilibrium phenomenon, an observation that plays a crucial role in the theory we develop here.

Third, banks will try to devise contracts that provide strong incentives not to default. This may lead to contingency contracts in which both the rate charged and the availability of credit at a later date depend on the borrower's previous performance. In conjunction with the specialized knowledge

mentioned above, this type of contract ties particular borrowers to particular lenders, that is, creates a "customer market" of the sort described by Arthur Okun. Thus, although the credit market is "competitive" in the usual sense (free entry, many buyers and sellers), lenders will view different borrowers as highly imperfect substitutes, and borrowers will have the same attitudes about different lenders—at least in the short run. There may, in particular, be classes of borrowers (like small businesses) for whom denial of credit by "their" bank has the effect of making credit inaccessible.

#### IV. The Effectiveness of Monetary Policy

We are now prepared to see how monetary policy affects real activity in this model. Consider what happens if the central bank sells bonds in the open market, causing a drain of reserves from the banking system.

If banks were essentially "loaned up" before, they will have to contract their loan supply. Some borrowers will not have their loans renewed. As we have just argued, many of these borrowers will be unable to secure credit from other banks. Investment activities will be curtailed and, if the loans were providing working capital, even current operations may have to be reduced. Thus tight money can depress real economic activity. Note also that, because of credit rationing, all this may happen with little increase in interest rates. So the effectiveness of monetary policy in this model does not rely on large interest elasticities, which often cannot be found empirically.

Two important questions remain. First, what stops prices from falling so fast that neither the real supply of credit nor real output has to decline, thereby robbing monetary policy of its real effects? Second, why do borrowers that are denied credit by the banks not turn elsewhere, for example, to the auction market?

The first question is as old as monetary theory itself, and bedevils any attempt to provide a deep explanation of the real effects of monetary policy. Part of the answer is simple and quite general: expected price changes affect the expected returns on hold-

ing financial assets (such as money), and therefore have real effects.<sup>2</sup> But we have just expressed doubts about the empirical importance of interest elasticities of this sort.

The rest of the answer has to do with the fact—the unexplained fact—that many long-term contracts without complete indexation exist. We do not have a good explanation for this phenomenon. Neither does anyone else.<sup>3</sup> But that does not imply that the consequences of nominal rigidities should be ignored. This paper seeks to explain how monetary policy works in the presence of such rigidities.

The second question is more specific to our approach. Recall that we rejected the transactions mechanism as an explanation for the real effects of money on the grounds that there were too many close substitutes. Analogously, our theory would not hold up if close substitutes for bank credit were readily available. Are there close substitutes?

If information were perfect (or cheaply acquired), then a reduction in bank credit would be offset by an increase in nonbank credit. Central bank policy would change the locus of borrowing, but would change neither the total volume of credit nor who gets it. However, we have argued that costly and specialized information is the essence of the credit market, so that good substitutes for bank credit do not exist, at least in the short run.

What about the market for commercial paper, for example? For some large firms (like General Motors) this is a real option, and they use it. For these firms, curtailments of bank credit may be offset by expansions of open market credit. But the fact of the matter is that for many firms, including all the small ones, commercial paper is simply not an option; if the banks are forced to

<sup>2</sup>Real effects can be avoided only by an exactly offsetting change in the nominal interest rate on financial assets. Naturally, this cannot occur in the case of currency.

<sup>3</sup>The analogy between the short-run rigidities imposed by multiperiod nominal wage contracts and those imposed by multiperiod nominal loan contracts should be apparent. For one attempt to explain why wages and interest rates may not be fully indexed, see Blinder (1977).

contract, they end up credit constrained. Thus, like Stiglitz and Andrew Weiss, we view the credit market as divided into clienteles. Very-low-risk borrowers can use the open market, and are rationed only by price. Very-high-risk borrowers cannot get credit at any price. Those in between may encounter quantity constraints, and this rationing becomes more severe when the central bank drains reserves from the banking system.

Notice that the segmentation of credit markets should become particularly severe during recessions, when even large, well-known firms face the possibility of default. Since investors assume that banks have superior knowledge about their customers, a firm that comes to the open market because it was rationed by its bank will be viewed as a bad risk, and therefore either charged a higher interest rate or denied access to the market.

Not much has been said so far about money; the emphasis has been on credit. To relate the two, consider a typical bank with equal liabilities (deposits,  $D$ , and net worth) equal to assets (reserves,  $R$ , loans,  $L$ , and government bonds,  $B$ ). Under a system of fraction reserve banking in which lending institutions also provide the medium of exchange (deposits),  $L$  and  $D$  will be closely related. Take our previous example in which the central bank makes an open market sale of government bonds:  $B$  rises and  $R$  falls by an equal amount. Banks then find themselves short on reserves and, as mentioned above, must contract  $L$ . But if  $R$  and  $D$  are held in fixed proportion, then the decline in deposits—and therefore in the money supply—must match the decline in loans.

Thus, while we have two competing theories—one based on credit, the other on money—that are conceptually distinct, the data will have difficulty distinguishing between them because credit and money normally are highly collinear. Given an institutional structure in which the same institutions supply loans and the medium of exchange, devising tests to distinguish between the “credit” theory and the “money” theory is no easy matter. And we do not pretend to have done this. However, we can make some suggestive remarks.

First, Benjamin Friedman has documented the facts that (a) a broad measure of credit (far broader than bank credit) does just as well as money in forecasting future movements in nominal *GNP*, and (b) credit is just about as closely related to the Fed’s instruments as is any of the monetary aggregates.

Second, Ben Bernanke’s study of detailed data from the Great Depression suggests that the decline in money was too small to account for the sharp drop in output, but that a proxy for credit stringency does rather well.

Third, the particular factors that have led to the breakdown of the demand function for money in recent years—deregulation and financial innovation—ought not to have destroyed the demand function for credit, according to the arguments presented here. In a period of rapid financial innovation, the ability of the central bank to curtail economic activity by causing a scarcity of the medium of exchange should be severely limited. Yet the Fed seems to have caused a severe disruption of economic activity, and has even done so without reducing the growth rate of money very much. We suggest that restrictions on the availability of credit, via the mechanisms discussed here, may provide a better explanation of how the Fed killed the economy.<sup>4</sup>

Finally, we should observe that, just as financial innovation has impaired the link between money and economic activity, further innovation might impair the link between bank credit and the economy. According to our arguments, it is the unique position of banks in the credit system that gives the central bank such strong leverage over the real economy. But if banks prove to be an unreliable source of funds, alternative institutions may arise that serve the same functions as banks. If such institutions do develop, the effectiveness of monetary policy might be seriously reduced.

<sup>4</sup>During the four years from 1978 through 1981, the December-to-December growth rate of what we currently call *M1* fell gradually from 8.3 to 6.4 percent, which hardly suggests a savage monetary squeeze. However, the growth rate of commercial bank loans fell from 18.1 to 6.4 percent.

## CLASSICAL ECONOMICS: THE SUBSISTENCE WAGE AND DEMAND-SUPPLY ANALYSIS

### Marx and the Iron Law of Wages

By WILLIAM J. BAUMOL\*

Marx den Laf [argue] sagte: Ce qu' il  
ya de certain c'est que moi je ne suis  
pas Marxiste.

*Letter from Engels to Bernstein*  
London 2/3, November 1882

I find few things as discouraging as the persistent attribution of positions to a writer whose works contain repeated, categorical, indeed emotional, denunciations of those views. Marx's views on wages are a prime example. Both vulgar Marxists and vulgar opponents of Marx have propounded two associated myths: that he believed wages under capitalism are inevitably driven near some physical subsistence level, and that he considered this to constitute robbery of the workers and a major evil of capitalism. Yet Marx and Engels tell us again and again, sometimes in most intemperate language, that these views are the very opposite of theirs. These observations, incidentally, are hardly new discoveries. Thus, for example, Roman Rosdolsky (1977, p. 287 ff.) disposes of the subsistence wage allegation and Robert Tucker (1969, ch. 3), and Allen Wood (1972) cover Marx's view on the morality of capitalist distribution very effectively.

#### I. Wages and Subsistence

I will show that to Marx the value of labor power does not normally equal physical subsistence; moreover, wages need not be equal to the value of labor power; with Marx's ascerbic rejection of the Malthusian model this system is denuded of *any* explicit wage equilibrium process; the omission of any fixed equilibrium point was deliberate, because Marx was anxious to show that workers

have the power to raise wages substantially even under capitalism; Marx considered the iron law of wages a monstrosity. These are not things he said once or twice, by indirection and in obscure places. They recur over and over, in *Capital* and in other writings including private notes and correspondence. Moreover, these are views Marx held to the end of his life.

#### A. Value of Labor Power and Physical Subsistence

A key statement in *Capital* does seem to support the subsistence allegation: the value of labor power is "...the labour-time necessary for the production and consequently also the reproduction, of this special article" (*Capital*, I, p. 189). But almost at once Marx points out that

...the number and extent of [the worker's] so-called necessary wants, as also the modes of satisfying them, are themselves the product of historical development, and depend therefore to a great extent on the degree of civilization of a country, more particularly on the conditions under which, and consequently on the habits and degree of comfort in which, the class of free labourers has been formed.

[*Capital*, I, p. 190]

This view was, of course, shared by Marx's predecessors (see, for example, Adam Smith, *Wealth*, p. 744; Ricardo [Sraffa, I, p. 97]). But it seemed particularly crucial to Marx. For example, he severely criticized the physiocrats (whom he generally admired) for maintaining that wages have a *fixed* floor (*Theories of Surplus Value*, I, p. 45). Similarly, in *Wages, Price and Profit*, to whose

\*Princeton University and New York University.

history I must return, Marx tells us

Besides [the] mere physical element, the value of labour is in every country determined by a *traditional standard of life*...the satisfaction of certain wants springing from the social conditions in which people are placed and reared up.... This historical or social element, entering into the value of labour, may be expanded, or contracted, or altogether extinguished....

[pp. 50–51, Marx's italics]

Other such quotations are easy to find. There simply can be no doubt about the matter. "The labour-time necessary for the production and...reproduction" of labor power is a flexible magnitude which is not "determined...by nature," that is, it is neither bare subsistence, nor any other preset amount.

#### B. Wages and Value of Labor-Power

Wages are, of course, the price of labor power. To most economists the "value of good  $x$ " connotes its price. But, to Marx, value clearly meant something else. His extensive discussion in Volume III of *Capital* of the transformation of values into prices deals with the persistent and systematic deviations between the two, even in equilibrium. Value was *defined* by Marx to equal a good's labor content (he repeatedly tells us it is a tautology) (see, for example, *A Contribution to the Critique*..., 1904, pp. 31–32), while Marx explicitly followed Smith and Ricardo in taking price to equal cost of production, including the normal return on capital (*Capital*, III, p. 233).

Now, since Marxian price differs from value, the wage rate, the price of labor power, like that of wheat, can and does differ from its value. It must be admitted that this distinction does not occur frequently in Marx's writings, but it certainly occurs explicitly. It is found, for example, in *Capital* (III, 1966 ed., p. 235) and in *Marginal Notes on Adolph Wagner*: "[Wagner claims I say] ...in the determination of the *value of labour power*, that its value is really paid, which is *not in fact the case*" (p. 43).

#### C. Absence of an Equilibrating Mechanism

Even though the Ricardians also did not believe that wages approach a fixed physical subsistence level, their model did have a mechanism driving wages toward the subsistence level *currently customary*. This mechanism was, of course, the Malthusian population principle which Marx rejected vehemently, for reasons I will discuss.

It has been suggested that the reserve army of the unemployed was substituted by Marx for the population principle. Marx does say that when wages rise, machinery is substituted for labor, and the resulting excess supply of labor limits these increases. But this only means that the slope of the labor demand curve is negative. There is clearly no way one can deduce from such a demand curve that the equilibrium wage must always (or ever) equal subsistence. We will see that Marx emphatically rejected such a conclusion, and why.

#### D. The Power of Workers over Wages

To understand the main piece of evidence on the degree to which Marx believed wages can be influenced by the workers, some biographical information is pertinent.

In 1865, two years before publication of Volume I of *Capital*, Marx was hard at work on the manuscript. Just then a member of the Council of the International Working Men's Association (John Weston, a follower of "Utopian" socialist Robert Owen) argued before the Council that unions can never raise real wages because wage increases must cause proportionate price increases. Marx pronounced this conclusion "theoretically false and practically dangerous" and undertook two lectures in reply to the Council. Marx wrote to Engels (May 20, 1865) that he did so most reluctantly both because it would give away some of *Capital*'s ideas and would take him away from his writing so that, to save time, his talks would be extemporaneous. It must have been a surprise when the manuscript (written in English) was discovered three decades later among Marx's papers by his daughter, Eleanor. The issue had evidently been most important to Marx.

In these talks, published as *Value, Price and Profit* (or *Wages, Price and Profit*) Marx was quite unambiguous on our subject:

By comparing the standard wages or values of labour in different countries, and by comparing them in different historical epochs of the same country, you will find that the *value of labour* itself is not a fixed but a variable magnitude, even supposing the values of all other commodities to remain constant...although we can fix the *minimum* of wages, we cannot fix their *maximum*.... It is evident that between the two limits...an immense scale of variations is possible. The fixation of its actual degree is only settled by the continuous struggle between capital and labour, the capitalist constantly tending to reduce wages to their physical minimum, and to extend the working day to its physical maximum, while the working man constantly presses in the opposite direction....

As to the *limits* of the *value of labour*, its actual settlement always depends upon supply and demand, I mean the demand for labour on the part of capital, and the supply of labour by the working men. [pp. 51–52]

Thus, Marx clearly believed that “an *immense* scale of variations” in wages is possible under capitalism and that it is the responsibility of unions to take full advantage of it. He emphatically did *not* believe that fate condemns workers to subsistence wages which they must accept passively. Note incidentally, that Marx believed unions to have a very valuable role, a piece of information which, one hopes, was some comfort to his daughter who had devoted much of her life to union activity, when she found the manuscript a few months before her suicide (see the superb biography of Eleanor Marx by Yvonne Kapp, 1972, 1976).

#### E. *The Iron Law of Wages*

A later manuscript, published as a *Critique of the Gotha Programme* (henceforth “*Gotha*”) can complete my discussion of Marx on wages and subsistence. Written in

1875 (and not intended for publication) it denounced a platform before a German socialist group meeting in the town of Gotha. The platform was an attempted compromise between Marxian principles and the sentimental notions of romantic socialist Ferdinand Lassalle, Marx’s ancient nemesis, then dead eleven years. (*Gotha* is the source of the phrase “from each according to his abilities, to each according to his needs”.) Several weeks before Marx wrote *Gotha*, Engels, undoubtedly with Marx’s knowledge, had written to one of the German Marxist leaders a letter that was to serve as an outline for much of *Gotha*. Engels wrote:

...our people have allowed the Lassallean “iron law of wages” to be foisted upon them...namely, that the worker receives on the average only the *minimum* in wages, and indeed because, according to Malthus’ theory of population, there are always too many workers.... Now Marx has proved in detail in *Capital* that the laws regulating wages...are in no sense iron but on the contrary very elastic.... The Malthusian argument in support of the law...has been refuted in detail by Marx in the section on the ‘Accumulation of Capital.’ Thus by adopting Lassalle’s “iron law” we commit ourselves to a false thesis with a false argument. [*Gotha*, Appendix, pp. 40–41]

Marx, in *Gotha*, denounces the Lassallean slogan: “the abolition of the wage system together with the iron law of wages,” writing:

...Lassalle’s attack on wage labour turns almost solely on this so-called law.... But if I take the law with Lassalle’s stamp on it and, consequently, in his sense, then I must also take it with his substantiation for it. And what is that?...it is the Malthusian theory of population.... But if this theory is correct, then again I *cannot* abolish the law even if I abolish wage labour a hundred times over, because the law then governs not only the system of wage labour but *every* social system. Basing themselves directly on this, the economists have been proving

for fifty years and more that socialism cannot abolish poverty, *which has its basis in nature*, but can only make it *general*, can only distribute it simultaneously over the whole surface of society! [Gotha, pp. 22–23]

Later I will show that this crucial passage helps to explain the reasons behind Marx's views. For now I use it just to confirm his abhorrence of anything like the "iron law" of wages.

## II. The Morality of Surplus Value

Marx repeatedly expresses contempt for the view that wages under capitalism are immoral and constitute grounds for revolution. In *Capital*, where he discusses the value of labor power, he describes as "a very cheap sort of sentimentality" the view that the method by which wages are determined under capitalism is "brutal" (Vol. I, p. 192). Similarly, about three years before he died, in *Marginal Notes on Adolph Wagner*, perhaps his last piece on economics, Marx remarked

[Adolph Wagner] foists on me the idea that "the *surplus-value* produced by the labourers *alone improperly* remains with the capitalist entrepreneurs" . . . In fact, I say the direct opposite: namely that at a certain point commodity production necessarily becomes 'capitalist' commodity production and that according to the *law of value* governing the latter, the "surplus-value" is necessarily the capitalist's and not the labourer's. [p. 61]

The same views appeared more than twenty years earlier, in the *Grundrisse*, and again in Engels' introduction to the first German edition of *The Poverty of Philosophy*:

According to the laws of bourgeois economics, the greatest part of the product does *not* belong to the workers who have produced it. If we now say: that is unjust, that ought not to be so, then that has nothing immediately to do with economics. We are merely saying that this economic fact is in contradiction

to our moral sentiment. Marx, therefore, never based his communist demands upon this...he says only that surplus value consists of unpaid labour, which is a simple fact. [pp. 10–11]

Marx himself returns to the morality of wages in *Capital*:

...The circumstance, that on the one hand the daily sustenance of labour-power costs only half a day's labour, while on the other hand the very same labour-power can work during a whole day...this circumstance is, without doubt, a piece of good luck for the buyer, *but by no means an injury to the seller*. [Vol. I, p. 216, emphasis added]

## III. Accounting for Marx's Positions

In sum, Marx consistently held views on the level, determination, and morality of wages very different from those attributed to him in popular legend. In his words, his positions are "the direct opposite" of those in the folklore. Why did Marx argue so vehemently that wages are not fixed at physical subsistence, and that the manner in which they are set is not immoral? We have rather clear indications of his own explanations.

First, as we have seen, Marx was anxious to encourage the activity of trade unions both "as centres of resistance against the encroachments of capital [upon wages... and] organized forces...for the final emancipation of the working class, that is to say, the ultimate abolition of the wages system" (*Wages, Price and Profit*, p. 55).

Second, the "iron law of wages" and the Malthusian model underlying it ascribe poverty not to any feature of capitalism but to human psychological propensities. As we have seen, Marx tells us in *Gotha* that this concedes everything to the opponents of socialism. For, if valid, it should be equally so in a socialist society which, consequently, could do nothing to eliminate poverty, aside from sharing the wealth or, rather (as Marx notes), sharing the poverty. No wonder Marx rejected the Malthusian model so vehemently, describing it as "a libel on the human race."

Third, Marx believed as a fundamental matter of philosophy that there is no such thing as an absolute standard of morality. A basic component of historical materialism is the view that no phenomena, and, emphatically, no social phenomena can be understood except in their historical context. Any proposition is robbed of its sense if it is taken as an eternal verity or as a truth independent of historical circumstances. Moral values (including value judgments on distribution) are no exception—a behavior pattern which is considered monstrous today may have to be adjudged ethical and appropriate for another society.

The idea of equality, therefore, both in its bourgeois and in its proletarian form, is itself a historical product, the creation of which required definite historical conditions which in turn themselves presuppose a long previous historical development. It is therefore anything but an eternal truth.

[Engels in collaboration with Marx, *Anti-Dühring*, p. 121]

This was one of the prime grounds on which Marx and Engels rejected the doctrines of the utopian and the romantic socialists. As Engels wrote:

To all these Socialism is the expression of absolute truth, reason and justice, and has only to be discovered to conquer all the world by virtue of its own power. And as absolute truth is independent of time, space, and of the historical development of man, it is a mere accident when and where it is discovered. With all this, absolute truth, reason, and justice are different with the founder of each different school... there is no other ending possible in this conflict of absolute truths than... a kind of eclectic, average Socialism... a mish-mash allowing of the most manifold shades of opinion; a mish-mash of such critical statements, economic theories, pictures of future society by the founders of different sects, as excite a minimum of opposition.

[*Socialism: Utopian and Scientific*, end of Section 1]

Given Marx's and Engels' fear of the vulnerability of the socialist movement to seduction by indefensible romantic notions which can undermine both its effectiveness and its purpose, their anxiety to hammer home such philosophical points is entirely understandable.

Finally, Marx was determined to battle the "iron law" and those who considered it a primary indictment of capitalism because it cheapens and trivializes the entire socialist cause.

It is as if, among slaves who have at last got behind the secret of slavery and broken out in rebellion, a slave still in thrall to obsolete notions were to inscribe on the programme of the rebellion: Slavery must be abolished because the feeding of slaves in the system of slavery cannot exceed a certain low maximum! Does not the mere fact that the representatives of our Party were capable of perpetrating such a monstrous attack on the widespread understanding among the mass of our Party prove by itself with what criminal levity and with what lack of conscience they set to work.... [*Gotha*, p. 24]

## REFERENCES

- Engels, Frederick, *Anti-Dühring*, London: Lawrence and Wishart, 1878, 1934.
- , *Socialism: Utopian and Scientific*, pamphlet extracted by Engels from *Anti-Dühring*, with a forward by Marx, Paris, 1880; Peking: Foreign Language Press, 1975.
- Kapp, Yvonne, *Eleanor Marx*, New York: Pantheon Press, 1972, 1976.
- Marx, Karl, *Capital*, Chicago: Charles H. Kerr, Vol. I, 1894, 1909; Moscow: Progress Publishers, 1966.
- , *Critique of the Gotha Programme*, 1875, Moscow: Progress Publishers, 1937.
- , *A Contribution to the Critique of Political Economy*, 1859, Chicago: Charles H. Kerr and Co., 1904.
- , *Grundrisse*, Martin Nicolaus, ed., trans., Harmondsworth: Penguin Books, 1973.
- , *Marginal Notes on Adolph Wagner's*

- '*Lehrbuch der politischen Ökonomie*', (1879-1880), in Marx-Engels, *Werke* (MEW) Vol. 19, pp. 355-383; Trans. by Athar Hussain in *Theoretical Practice*, Issue 5, Spring 1972.
- \_\_\_\_\_, *The Poverty of Philosophy*, London: Martin Lawrence [1846-1847], N.D.
- \_\_\_\_\_, *Theories of Surplus Value*, Moscow: Progress Publishers, 1963.
- \_\_\_\_\_, *Value, Price and Profit*, New York: International Publishers, 1898, 1935.
- Ricardo, David, *Principles of Political Economy and Taxation*, P. Sraffa, ed., Vol. I, *The Works and Correspondence of David Ricardo*, Cambridge: Cambridge University Press, 1951.
- Rosdolsky, Roman, *The Making of Marx's 'Capital'*, London: Pluto Press, 1977.
- Smith, Adam, *The Wealth of Nations*, Cannan ed. Modern Library, New York, 1937.
- Tucker, Robert, *The Marxian Revolutionary Idea*, New York: W. W. Norton & Co., 1969.
- Wood, Allen, "The Marxian Critique of Justice," in *Philosophy and Public Affairs*, Vol. 2, Spring 1972.

# The Classical Theory of Wages and the Role of Demand Schedules in the Determination of Relative Prices

By PIERANGELO GAREGNANI\*

My purpose in this paper will be twofold. First, I shall argue that the role of demand functions in determining prices depends on their role in determining distribution by means of the "relative scarcity" of the "factors of production." As a result, these functions would have no role in determining prices in the approach of Adam Smith and Ricardo who did not explain distribution in that way. These considerations will lay the ground for my second purpose: to distinguish between the notion of demand schedules for commodities and that of "effectual demand" in Smith and Ricardo, and to contend that the attempt to read in the classical authors as explanation of relative prices along the lines of modern theory is not well founded.

## I. Demand and "Marginalist" Prices

The notion of demand schedule requires that the price-quantity relationship be determinate for all prices in the relevant range, and not only for the "natural" or "normal" price, which, however, is the only one that we may expect to experience under the non-accidental conditions that are likely to emerge through a repetition of the situation. We are therefore dealing with a much stricter notion than the immediately plausible one according to which an accidental fall in the quantity supplied below its normal level is likely to be accompanied by a rise in the price, and vice versa: in this notion no attempt would be made to determine the magnitude of such a rise, considered as depending on accidental factors.

This second, weaker notion (which, as I shall contend below, is that held by the classical authors) could not be represented

by a curve in the familiar diagram: the prices corresponding to quantities below (above) the normal quantity  $q_n$  would be determinate only in that they are higher (lower) than the normal price  $p_n$ . If we wished to represent this notion in such a diagram, we would find *two areas*, North-West (*NW*) and South-East (*SE*) of the normal price-quantity point  $P$ , where *NW* indicates where the price is likely to be found when the quantity supplied has fallen accidentally short of  $q_n$ , and *SE* indicates where it is likely to be in the opposite case. To pass from this diagram to the familiar demand curve requires the assumption that the price-quantity relations falling into those two areas are as definite as they are at the normal point  $P$ . This, though formally tempting, cannot be done without a theory which allows us to determine those points.

The theory which has been advanced to that effect is the dominant one in its two aspects of: (i) asserting definite tastes for each consumer such that, given his income and any set of relative prices, the quantities of goods he demands are determined; (ii) ensuring individual income levels corresponding to the full employment of their productive services or, more generally, determinable simultaneously with the demand price of the commodity and undergoing comparatively small changes as the quantity supplied changes. (The demand function is based here on the general equilibrium system, but any "partial equilibrium" notion of it rests on its general equilibrium counterpart to which we should refer in order to ascertain its properties and adequacy.)

The same analysis ensures a persistence of the demand function sufficient to correct accidental deviations from it through repetition over time. Such persistence will in fact be that of individual tastes and of the other data of the system.

\*University of Rome.

It is, on the other hand, generally recognized that the same theory allows us to conclude that the demand functions will generally show a negative price-quantity relation, and thus to argue for the uniqueness and stability of the equilibrium concerned. This negative relation, I should note, requires a specific ordering between *each* price-quantity point. Even more importantly, the theory does not regard these points as results of accidental and temporary deviations of the quantity supplied from the "normal" level, but rather as determinate points likely to emerge from a repetition of the event.

If the notion of the demand function for a product depends upon the dominant theories for its rational basis, its role in determining the relative prices of the products depends on the idea, which is characteristic of these theories, that the distribution of the social product is determined by an "equilibrium" between the "demand and supply" of the services of factors of production.

In fact, let us assume, as is generally done in these theories, that constant returns to scale prevail in each industry. Let us also assume, at first, that land is free. It follows that, given the real wage, or, alternatively, the rate of profit (interest), the relative prices of all products will be determined *independently* of any demand functions. Thus the demand functions can enter into the determination of the prices of products only to the extent to which they enter into the determination of the division of the product between wages and profits. And the demand functions for products do enter the determination of distribution in modern theory, because the decreasing demand functions for services of factors are derived from them, as well as from the conditions of technical substitution.

This conclusion, according to which a demand schedule can only affect the price of the corresponding product to the extent to which it affects distribution, might at first puzzle the reader used to thinking in terms of the intersection of the demand and supply curves of the product in question. It is evident, however, that the demand curve of a product can affect its price only if the supply curve is nonhorizontal. Now, under the as-

sumption of constant returns to scale, the supply curve will have a rising slope because the expansion of output will generally render the factors used in relatively higher proportions for the product in question more costly. Indeed the nonhorizontality of the supply curve is the *expression* of the extent to which the quantity produced, and hence the demand conditions of the commodity, affect distribution. The same nonhorizontality is, on the other hand, the *agency* through which this effect on distribution makes itself felt on the price of the commodity.

This role of consumer choice in determining prices seems, at times, to be imperfectly grasped in the literature. An example is perhaps provided by the sense of novelty that the "nonsubstitution theorems" have aroused. Thus, to take the most significant of these theorems, Paul Samuelson (1961, p. 528) found that in an economy where production requires only labor and capital goods, "a stipulated change in the pattern of demand for end-goods" will affect neither the relative prices of such goods nor the methods of production in use, once the rate of interest (profit) is given. This proposition would have seemed less novel had it been clear that the pattern of demand can only affect relative prices.

## II. Demand and "Classical" Prices

When this role of demand functions in the determination of the prices of products is understood, it should become clear why they are not to be found in Adam Smith and Ricardo, who had a different theory of distribution. These authors envisaged the real wage as dependent, essentially, on institutional factors, together with the conditions affecting what might perhaps be summed up as the relative bargaining position of workers and employers.

Thus Adam Smith attributed a central role to the notion of a culturally determined level of subsistence (*Wealth*, II, pp. 351-52) and held that the tendency towards such a level was largely explained by the "advantage" which masters have in disputes over wages, both because they are always "in a sort of tacit, but constant and uniform combination,

not to raise the wages of labour" and because in case of dispute they can "hold out much longer" (*Wealth*, I, p. 59). Ricardo, for his part, also focused his explanation of the real wage on a notion of subsistence which, determined by historical no less than by biological factors, was in the given conditions the minimum acceptable by workers for any length of time. For the tendency of the "market" wage towards such a "natural" level, Ricardo, influenced by Malthus, relied on changes in population, which Smith had also considered, though more flexibly than Ricardo. Ricardo's own position was however far from rigid: he freely admitted that "with better education and improved habits" the natural wage could itself rise (compare, for example, Ricardo's spirited reaction to Malthus in *Works*, II, p. 115).

It seems therefore that what characterized these authors was not the idea of a wage determined by subsistence, even less that of a subsistence constant over time. It was, more generally, the importance attributed, in the determination of the real wage, to elements which were best studied before and independently of the determination of relative prices and of the other shares in total product. This *separate determination* found expression in the fact that these authors took the real wage as given when approaching the determination of relative prices. This in turn implied that the price system and the rate of profit could be determined independently of any demand functions for the products.

An interpretation of Adam Smith and Ricardo as "modern economists trying to be born," holding, that is, a demand and supply analysis of wages while seeking to say something "significant and limiting about their properties" (p. 1415), has been advanced by Samuelson (1978) with his "canonical classical model." According to it, Smith and Ricardo would have determined the "equilibrium" real wage as that balancing the growth of the supply of labor with that of its demand, resulting from accumulation (p. 1416). The "demand" for labor of this interpretation would be rigid, implying, as Samuelson notices (p. 1423), either zero wages or zero profits or an indeterminate breakdown between wages and profits—

which would be limiting indeed for "modern economists." But, above all, this interpretation and, particularly, the relation between the real wage and growth of population on which it is based, seems to suffer from the tendency to see functional relations of known general properties where the classical economists saw relations too complex and variable to be quantified in any exact way. Thus Smith wrote that "the liberal reward of labor," by enabling workers to provide for their children, tends to increase population (*Wealth*, I, p. 71), but he also brought out elements which went in the opposite direction (see, for example, *Wealth*, II, p. 353), or could go in either direction (*Wealth*, I, pp. 62–63) (on this see also Joseph Spengler, 1959, p. 7). I have already mentioned the flexibility of Ricardo's position on this matter.

It might be objected that my argument concerning the classical economists has so far rested on the assumption of constant returns to scale to labor and capital and that, when this assumption is abandoned, a second route emerges through which outputs and hence, presumably, demand functions, may affect prices even when the real wage is given.

What this objection presumes is that demand functions must be introduced to determine outputs even when distribution is otherwise determined. We may here lay aside the difficulties of envisaging these functions in a classical setting (compare my earlier discussion on the determination of consumer incomes). It is the usefulness of this procedure which is questionable in the first place. As noted above, the modern analysis of demand is in fact mainly concerned with some *formal properties* of consumer tastes, specifically with the determinateness, persistence, and slope of the demand curves, and *not* with the actual *content* of these tastes (which is generally left to the sociologist or psychologist). This content, jointly with the levels of activity, distribution and techniques, is however, what determines the *position* of the demand curve and is thus the main influence on the levels of output. Now, those formal properties, basic as they are for the modern supply-and-demand analysis of distribution,

were largely irrelevant for the classical economists with their different theory. It was therefore natural that these authors should, so to speak, face the content of consumer tastes *directly*, without the intermediate screen of any formal properties, whether in order to take it as given (as is generally done in modern theory) or in order to examine it, as they generally did (for example, in connection with workers' "necessaries"). In either case, the procedure was to take this content as given *when determining* the system of relative prices—leaving it for a separate analysis, like that to be conducted for the other determinants of output (real wages, levels of activity, and techniques). The very levels of output could then be taken as given in just the same way as the real wage was taken as given.

The analysis of changes in outputs and prices was thus conducted by the classical economists in what we may describe as two distinct logical stages: (i) the effect on relative prices of the change in real wages, or techniques, or outputs was examined with outputs as independent variables; (ii) the possible effects on outputs of the change in relative prices were then analyzed in accordance with the circumstances of the case under consideration, jointly with any possible further effects on prices and distribution due to nonconstant returns to scale.<sup>1</sup> With this the classical economists distinguished between field of analysis (i), where necessary quantitative relations could be found between rates of remuneration, and between these rates and relative prices, and other fields where no such necessary relations could be established, and where actual relations had to be studied in their multiplicity and diversity according to circumstances. This procedure by separate logical stages is in fact nothing new, even for modern theory: it will, for example, be generally admitted that technical changes will generally affect consumers' tastes, but any such effect will be considered, if at all, at a stage which is logically distinct

from the determination of distribution and prices, where consumers' tastes and technical conditions of production appear as data. It remains, however, true that the procedure of the classical economists renounces what was attempted by later theory, namely, a simultaneous treatment of the interrelations between most economic phenomena. This modesty of goals may however be the most appropriate one in a subject as complex as economics where, as Marshall reminds us, "the function... of analysis and deduction... is not to forge a few long chains of reasoning, but to forge rightly many short chains" (*Principles*, Appendix C; 3; p. 773).

### III. The Classical Notion of "Effectual Demand"

In Adam Smith and Ricardo we find the notion of "effectual demand," defined as "the demand of those who are willing to pay the natural price of the commodity" (*Wealth*, Bk. I, ch. VII, I, p. 49). The analysis above should make it easy to see the difference between this notion and that of demand schedule. The role of effectual demand is to explain the tendency of the actual or "market" price toward the natural price and not that of determining the latter. It does not therefore consist of a curve but of a single determinate price-quantity point. Apart from this single point, Smith needs only to suppose that when the quantity supplied falls short of effectual demand, the actual or market price will exceed the natural price, thus setting in motion forces which tend to raise the quantity supplied and bring the market price down to the natural level (and vice versa).

This classical notion of demand, which was consistent with the theoretical framework of which it was an integral part, has however been often envisaged as a rudimentary expression of the modern notion of demand function. Marshall showed the way. In his *Principles* he refers to the "market" price—defined by Smith and Ricardo as the *actual* price, accidental in its absolute level and determinate only in its *order* relative to the natural price—as a "temporary equilibrium" price (see, for example, Marshall, *Principles*, Bk. V, ch. V, 8; I, pp. 378–79; also p. VII). With this he attributed to Smith

<sup>1</sup>As for the dependence of prices on outputs in the case of jointly produced commodities, the relative scarcity of these commodities will tend to find an objective expression in the co-existence of processes producing the same commodities (compare Sraffa, p. 43).

a demand curve, determinate also at points (like that of the "equilibrium" market price) other than the effectual demand point. After this first step, it was easy for Marshall to proceed and argue that the distinction between market and normal price was only one of degree, relating to the period of time over which the equilibrating process was supposed to occur, thus implicitly attributing to Smith and Ricardo also a demand-and-supply determination of the normal price (on this point, see Krishna Bharadwaj, pp. 264-65).

If, in Smith's case, the shortcomings of this interpretation could pass almost unnoticed, it was inevitable that they should crop up in the case of the more consistent version of the theory provided by Ricardo. In particular, it seemed clear that Ricardo determined prices independently of anything resembling demand schedules.

As is well known, Marshall attempted to cope with this difficulty by contending that Ricardo could avoid referring to demand and utility only by assuming what Marshall characterized as the "law of constant return," by which in the first place he meant a horizontal long-period supply curve for the industry in question (*Principles*, p. 671). Marshall here overlooks or leaves aside the fact that the dependence of prices on demand functions is based on distribution and not on the laws of returns to scale to capital and labor—to which he might also be taken to refer with his "laws of return." Indeed Ricardo could have deduced a constant supply price from constant returns to scale to capital and labor only to the extent that he was not explaining the distribution of wages and profits in Marshall's way. The incorrectness of Marshall's interpretation is, on the other hand, evident from the fact that in agriculture, where clearly Ricardo did not assume "constant return" in either sense, he did not introduce demand functions any more than he had done for manufacturing.

Gerald Shove took up this argument, but was more cautious than Marshall in attributing demand functions to Ricardo. He perceived that a "step" was involved in passing from Ricardo's discussion of the "market"

price to a demand function (p. 301) and saw a dilemma in Ricardo's theory of value: either introducing demand functions or remaining confined to the special case of constant supply price (p. 297). This unreal dilemma has recently been revived by S. C. Rankin who claims, p. 251, that, in Ricardo, *demand functions* do enter the determination of prices: his evidence is, however, that Ricardo has agricultural prices dependent on "demand," by which Ricardo meant "effectual demand," as is clear in the passages quoted by Rankin.

## REFERENCES

- Bharadwaj, Krishna, "The Subversion of Classical Analysis: Alfred Marshall's Early Writing on Value," *Cambridge Journal of Economics*, September 1978, 2, 253-71.
- Marshall, Alfred, *Principles of Economics*, London: McMillan & Co., 1920.
- Rankin, S. C., "Supply and Demand in Ricardian Price Theory: a Re-Interpretation," *Oxford Economic Papers*, July 1980, 32, 241-62.
- Ricardo, David, *The Works and Correspondence of David Ricardo*, Sraffa ed., Vol. I-IX, Cambridge: Cambridge University Press, 1951-58.
- Samuelson, Paul, A., "A New Theorem on Non-Substitution," in Ugo Hegeland, ed., *Money, Growth and Methodology and Other Essays in Economics; In Honor of Johan Akerman*, Lund: CWK Gleerup, March 1961.
- \_\_\_\_\_, "The Canonical Classical Model of Political Economy," *Journal of Economic Literature*, December 1978, 16, 1415-34.
- Shove, Gerald F., "The Place of Marshall's *Principles* in the Development of Economic Theory," *Economic Journal*, December 1942, 52, 294-329.
- Smith, Adam, *The Wealth of Nations*, London: J. M. Dent & Sons, 1950.
- Spengler, Joseph, "Adam Smith's Theory of Economic Growth," Part II, *Southern Economic Journal*, July 1959, 26, 1-12.
- Sraffa, Piero, *Production of Commodities by Means of Commodities*, Cambridge: Cambridge University Press, 1960.

# On the Interpretation of Ricardian Economics: The Assumption Regarding Wages

By SAMUEL HOLLANDER\*

Some thirty years ago, George Stigler published what was to become a standard reference for students of the classical period—"The Ricardian Theory of Value and Distribution." A key feature of the theoretical system developed in the *Principles of Political Economy* is there said to be the *subsistence wage theory*, which (together with the measure of value) was added to those elements already present in the earlier *Essay on Profits*, namely, the theory of rent and the dominant influence of diminishing returns in agriculture upon the rate of profit (1965, p. 187). From these elements there followed the "great conclusion" of the model: "With the growth of population, the rate of wages rises" (reflecting the increasing real cost of producing the *given* basket), "the rate of profit falls, and aggregate rents rise—all in terms of the measure of value" (p. 190).

In a review of my recent study of David Ricardo (1979), Stigler reiterates by implication this evaluation of the nature of Ricardo's contribution and explicitly repeats the attribution to Ricardo of the subsistence wage assumption. It remains his belief that the logic of Ricardo's argument requires the assumption: "I am amazed to be told that 'all the evidence' points to Ricardo believing that the wage will fall secularly"; similarly, "without this assumption [Ricardo's] fundamental theorem on distribution (only a rise in wages will lower profits) is not rigorously true"; and "his chapter on gross and net revenue and his repeated proposition that only net revenue (which excludes wages) can be taxed or saved are wrong" (1981, p. 101).

Much depends on the precise assumption made regarding wages; the subsistence wage attribution, for example, has distorted the entire body of "Cambridge" historiographical doctrine relating to classical economics.

Stigler is right to focus on this issue. But I believe it can be shown that his version of Ricardianism is invalid. If we are ever to do justice to the historical Ricardo and to the course of nineteenth-century economics we must by all means abandon the fix-wage attribution.

## I

I shall proceed by drawing upon another well-known contribution by Stigler. In his paper on "Textual Exegesis as a Scientific Problem" (1965), he observes that in seeking an accurate interpretation "we should not be so literal-minded as to count the passages in a book to decide an author's general position because the passages are not of equal importance" (p. 448). Where there exists a clash of texts the solution is to investigate the dependency of a writer's main analytical conclusions upon the alternative readings: "We increase our confidence in the interpretation of an author by increasing the number of his main theoretical conclusions which we can deduce from (our interpretation of) his analytical system" (p. 448). This procedure Stigler refers to as "*scientific exegesis*," and is designed to isolate the "net scientific contribution" of the writer. If, however, the historian is concerned with what the author in question "really believed," then the criterion should be which of the various alternative interpretations best suits what is known of the man's style—"personal exegesis."

Let us accept that the inverse wage-profit relation and the falling secular profit rate are amongst Ricardo's "main analytical conclusions." How does the notion of scientific exegesis help us? Were Stigler correct that these results require the constant wage assumption, the answer would be self-evident. But he is mistaken. A model can be devised, incorporating the fundamental theorem and

\*University of Toronto.

generating the falling return on capital, wherein the wage rate is a *variable*. The model in question is a genuine growth model in the sense that until stationarity is achieved, the wage rate and the profit rate are both *above* their respective "minima" encouraging net capital accumulation and population growth; and both will ultimately decline from whatever their "present" values may be to those respective minima in consequence of the pressures increasingly exercised by land scarcity. Only in the stationary state itself is the wage at subsistence and the profit rate at its minimum. Versions of this model were developed independently by inter alia John Hicks and myself (1977), Carlo Casarosa (1978), and Paul Samuelson (1978).

Not only is a variable-wage growth model technically meaningful, it is precisely what Ricardo himself specified to be his own. Ricardo explicitly describes the *falling* commodity wage of the secular path:

In the natural advance of society, the wages of labour will have a tendency to fall, as far as they are regulated by supply and demand; for the supply of labourers will continue to increase at the same rate, whilst the demand for them will increase at a slower rate. If, for instance, wages were regulated by a yearly increase of capital, at the rate of 2 per cent., they would fall when it accumulated only at the rate of 1-1/2 per cent. They would fall still lower when it increased only at the rate of 1, or 1/2 per cent., and would continue to do so until the capital became stationary, when wages also would become stationary, and be only sufficient to keep up the numbers of the actual population. [1951, Vol. I, p. 101]

The complication deriving from the rising prices of wage goods is then allowed for; although the commodity wage falls the money wage will rise:

As population increases, these necessities will be constantly rising in price, because more labour will be necessary to produce them. If, then, the money wages of labour should fall, whilst every commodity on which the wages of

labour were expended rose, the labourer would be doubly affected, and would be soon totally deprived of subsistence. Instead, therefore, of the money wages of labour falling, they would rise; but they would not rise sufficiently to enable the labourer to purchase as many comforts and necessities as he did before the rise in the price of those commodities. [1951, Vol. I, p. 101]

It is precisely this rise in the money wage that causes the profit rate to fall despite the decline in the commodity wage.

The model can also easily incorporate an initial section of rising commodity wages; and Ricardo thought that it should (see my 1979 book, p. 395ff.; also my paper with Hicks, p. 365). Stigler is unaware of the *declining* path of wages, but in 1952 he did allude to Ricardo's references to *rising* wages. But these, he says, "must simply be recorded as correct views which Ricardo did not know how to incorporate into his theoretical system" (1965, p. 172).

Poetic justice indeed! To attribute the fix-wage model to Ricardo is illegitimate in terms of Stigler's own criterion of scientific exegesis. For what Ricardo insisted upon was a secular fall in the profit rate *along with* an initial stage of increasing wages and a final stage of declining wages and Stigler's version simply cannot accommodate this complexity.

I do not intend to suggest that Ricardo's account is faultless. It would be remarkable were that the case. In fact, it is probable that there are two Ricardo models—those characterised by the versions offered by Hicks and myself, and by Samuelson or Casarosa. On the first view, secular variations in real wages are the outcome of *differential* growth rates of capital and labour—the upward trend a result of capital growth outpacing labour growth at an early stage of development and the downward trend a result of the reverse relationship (see the citation above from Ricardo). This version does not, in short, portray a "dynamic equilibrium" or balanced-growth path of wages.

Such a path is expounded in Samuelson's "canonical classical model." It traces out the unique values of the wage, given the general

labor and capital supply functions, which assure balanced factor growth subject always to the constraint imposed by land scarcity; the wage falls although labor supply decelerates in line with capital. There is much evidence to suggest that this version too is a valid attribution to Ricardo (see Section IV below), although the clearest exposition, which Ricardo accepted, is by Malthus (1820, see Ricardo, Vol. II, pp. 255–56).

## II

What is the source of Stigler's misunderstanding? His own notion of personal exegesis provides the key.

That Ricardo's *Principles* and his *Essay* are replete with references to the constant wages assumption has never been denied. The growth model in the *Essay on Profits* is expounded on the basis of a constant wage though not at the subsistence level (see my 1979 book, p. 136). But Ricardo himself tells us precisely why he so proceeds; he was no methodological neophyte:

We will, however, suppose that no improvements take place in agriculture, and that capital and population advance in the proper proportion, so that the real wages of labour continue uniformly the same;—that we may know what peculiar effects are to be ascribed to the growth of capital, the increase of population, and the extension of cultivation, to the more remote, and less fertile lands [1951, Vol. IV, p. 12]

All that he intended was a simplifying assumption for the sake of clear exposition—to allow him to focus upon one causal variable at a time playing on profits. There is no reason to believe that anything more was at stake when he devised the famous numerical illustration of the *Principles* (chs. 5; 6) which incorporates the constant wage assumption. Stigler has, it seems, mistaken Ricardo's typical "first approximations" for a full growth model—the man's *style* for the substance.

## III

The same conclusion follows in the wage taxation context. As I explain in my book

(1979, p. 383) Ricardo's taxation theorems can be interpreted as applying to the case where a subsistence wage rules. I provide textual evidence from the chapter "On Wages" and elsewhere to show this. But I then demonstrate that the wage taxation theorems do not stand or fall with the subsistence assumption (1979, p. 386). Stigler is mistaken when he cites passages in Ricardo (1951, Vol. I, pp. 215; 219) purportedly indicating adherence to a *subsistence* wage, for it is a constant wage *above subsistence* to which Ricardo there alluded. A word of explanation.

During the course of his discussion (in the chapter on "Taxes on Raw Produce") of the effect upon the money-wage rate induced by the taxation of necessities, Ricardo introduced the qualification that the "rate of progression" of the economy is throughout taken for granted, clearly implying that the analysis was intended to apply whether or not wages are initially at "subsistence":

Those who maintain that it is the price of necessities which regulates the price of labour, *always allowing for the particular state of progression in which the society may be*, seem to have conceded too readily, that a rise or fall in the price of necessities will be very slowly succeeded by a rise or fall of wages. [1951, Vol. I, p. 161, emphasis added]

If this statement were the only one of its kind, it might perhaps be dismissed as unrepresentative. But the fact is that the chapter "On Taxation of Wages" itself is formally contingent upon it.

The chapter "On Taxation of Wages" unlike the one "On Wages" is based upon Adam Smith's proposition that "the demand for labour, according as it happens to be either increasing, stationary, or declining, or to require an increasing, stationary, or declining population, regulates the subsistence of the labourer, and determines in what degree it shall be either liberal, moderate, or scanty" (1951, Vol. I, p. 215). And the general applicability of the taxation theorems is much emphasized:

"The price of labour will express, clearly, the wants of the society respec-

ting population" [Malthus]; it will be just sufficient to support the population, which at that time the state of the funds for the maintenance of labourers, requires.... Suppose the circumstances of the country to be such, that the lowest labourers are not only called upon to continue their race, but to increase it; their wages would be regulated accordingly. Can they multiply in the degree required, if a tax has taken from them a part of their wages, and reduces them to bare necessities?

[1951, Vol. I, pp. 219–20]

That Ricardo should have proceeded in the chapter "On Wages" and in other contexts on the assumption of a subsistence wage despite his acceptance of the Smithian position according to which the equilibrium wage is that wage which assures an appropriate positive rather than a zero growth rate of labor supply, is not difficult to appreciate. It is characteristic of his general method to simplify the analysis wherever this can be done without loss. The broader application of the taxation theorems means simply that the mechanism of population adjustment to wages above or below a given "subsistence" rate must now be applied to wage variations about a long-run labor supply curve relating the wage rate to the growth rate of population. Once again Stigler has mistaken a strong-case simplifying assumption for the entire analysis.

#### IV

Ricardo's full position as formulated in the taxation context—that of Smith and Malthus—has broad implications for growth in general. The rate of capital accumulation acts, in the first instance, as an independent variable which determines what the commodity-wage rate must be to guarantee an equivalent growth rate of population. Any disturbance which raises the price of wage goods—the effects say of increasing land scarcity and not merely taxation—necessitates an appropriate monetary compensation to assure that the growth rate of population is not impeded. The "natural" wage—albeit "unnatural" because a limiting case (1951, Vol. II, pp. 227–28)—is then the wage which

assures a constant population and will correspond to zero net accumulation, while to each positive growth rate of capital there will correspond a higher real-wage rate to assure the equivalent growth rate of labor supply. (A "high" but constant, rate of accumulation according to this analysis entails a high real wage rate; while an increase (decrease) in the rate of accumulation generates an increase (decrease) in the real wage rate.) But the assumption that the rate of capital accumulation is an independent variable—totally unaffected by the reduction in profits corresponding to the increase in money wages—is, however, no more than a first approximation. A reduction in profits would, it is conceded, probably have some effect on accumulation so that the compensatory increase in money wages would not entirely prevent a fall in real wages in consequence of taxation (1951, Vol. I, pp. 221–22; 225–26). In effect, the "dynamic equilibrium" path of the system has been altered to entail reduced (net) factor returns and factor growth rates.

The full analysis of wage taxation thus allows for a decline in the real wage and flies in the face of a fix-wage interpretation. All the ingredients of the canonical growth model—and as remarked earlier Ricardo approved of that version formulated by Malthus—come into play in the wage taxation context.

#### V

There remains Ricardo's conception of net revenue to which Stigler also alludes in support of his case. Malthus had made the same erroneous attribution in his *Principles* (1951, Vol. II, p. 381). His position, Ricardo complained, had been misunderstood by Malthus (and by Say). The inclusion, within net income, of profit and rent alone was merely a simplifying assumption without substantive intent:

Mr. Malthus says 'the additional two millions of men would some of them unquestionably have a part of their wages disposable.' Then they would have a part of the neat revenue. I do not deny that wages may be such as to give to the labourers a part of the neat

revenue—I limited my proposition to the case when wages were too low to afford him any surplus beyond absolute necessities

[1951, Vol. II, pp. 380–81]

Similarly regarding profits and rent as the source of accumulation and taxation:

Perhaps this is expressed too strongly, as more is generally allotted to the labourer under the name of wages, than the absolutely necessary expenses of production. In that case a part of the net produce of the country is received by the labourer, and may be saved or expended by him; or it may enable him to contribute to the defence of the country [1951, Vol. I, p. 348 fn.]

## VI

In another of his seminal articles (1976) Stigler makes the point that if we seek to understand the *scientific* role played by a figure in the history of economic theory then what is relevant is how his work “appeared to his contemporaries,” for science “consists of the arguments and the evidence that lead *other* men to accept or reject scientific views” (p. 60). The ideas they may have intended to express are not relevant from this perspective. This is very true and it is a perspective that casts much light on a number of issues in scientific history. But we must also allow for *erroneous* interpretations. While the variable-wage model was an open book—the above account is not based on correspondence or other unpublished and not easily accessible items—from the very outset readers have often misunderstood Ricardo. Erroneous

readings bedevil the problem of his legacy.

## REFERENCES

- Casarosa, Carlo, “A New Formulation of the Ricardian System,” *Oxford Economic Papers*, March 1978, 30, 38–63.
- Hicks, John and Hollander, S., “Mr. Ricardo and the Moderns,” *Quarterly Journal of Economics*, August 1977, 91, 351–69.
- Hollander, Samuel, *The Economics of David Ricardo*, Toronto: University of Toronto Press, 1979.
- Ricardo, David, *The Works and Correspondence of David Ricardo*, Vol. I, *Principles of Political Economy*, ed. Piero Sraffa, Cambridge: Cambridge University Press, 1821, 1951; Vol. II, *Notes on Malthus's Principles of Political Economy*, 1820, 1951; Vol. IV, *Pamphlets, 1815–1821*, 1951.
- Samuelson, Paul A., “The Canonical Classical Model of Political Economy,” *Journal of Economic Literature*, December 1978, 16, 1415–34.
- Stigler, George J., “The Ricardian Theory of Value and Distribution,” *Journal of Political Economy*, June 1952, reprinted in *Essays in the History of Economics*, Chicago: University of Chicago Press, 1965, 165–98.
- , “Textual Exegesis as a Scientific Problem,” *Economica*, November 1965, 32, 447–50.
- , “The Scientific Uses of Scientific Biography, with Special Reference to J. S. Mill,” in J. A. Robson and M. Laine, eds., *James and John Stuart Mill: Papers on the Centenary Conference*, Toronto: University of Toronto Press, 1976, 54–66.
- , “Review,” *Journal of Economic Literature*, March 1981, 19, 100–02.

## CHINESE ECONOMIC REFORMS

### Price Adjustment, the Responsibility System, and Agricultural Productivity

By THOMAS B. WIENS\*

The Government of the People's Republic of China considers that the prospects for rapid and broadbased economic growth in China depend crucially on the ability to increase production in the agricultural sector. The major instruments available to the government include reform of the institutional structure of farm management, modification of the farm price structure, and increase of budgetary expenditure in support of agriculture. Since 1978, the central government has stressed use of the first two instruments and some initial judgments can be made about the effectiveness of its measures based on recent official statistics.

#### I. Rural Institutional Reforms

Broad changes in rural policy beginning in 1978 include the attempt to eliminate excessively egalitarian practices, "unreasonable burdens and arbitrary exactions," and the use of coercion and regulation to control peasant farming practices. These reforms have been implemented with considerable thoroughness. Local officials seem committed to the improvement of farm incomes, and much more scope for private activity exists: personal income derived from collectively owned farms and other enterprises now represents only 52 percent of total rural income, while 38 percent is derived from private household activities, mostly from animal raising and retained land (private plots). Development plans emphasize diversification into oilseeds and cotton, animal husbandry,

sericulture, orchards and trees, and agro-processing, partly at the expense of grain area.

Dramatic changes occurred in 1980-81 as part of the implementation of various forms of the "responsibility system" (*RS*). The *RS* encompasses a variety of systems to organize management and distribution in the rural economy. These systems all lodge responsibility for day-to-day management of farming and subsidiary enterprises in units smaller than the production team (i.e., work groups, households, or individual workers). Contracts are utilized to precisely define responsibilities and attempt to link productivity with material reward. Although subdivision of ownership rights is not an acceptable element, complete subdivision of land management to the family level and abandonment of the systems of workpoint accounting which previously distinguished the Chinese commune are permissible. Many of the new organizational arrangements were worked out among farmers, often initially without knowledge or approval from higher administrative levels. However, by the end of 1981 official policy accepted even the extreme forms, as long as the above and certain other criteria were met; the systems were acceptable to participants; and the most decentralized forms were confined to areas which were poor and lacking in collectively owned capital. By mid-1981, in 28 percent of teams, management and often distribution decisions were being made at the household level (Tang Tsou et al.). While the prevalence of various systems was still in flux (generally further devolving) through the end of 1981, present policy is to stabilize the situation.

The *RS* is intended to provide the material incentives lacking under prior forms of

\*Agricultural economist, The World Bank. This paper is based on research completed before I joined the Bank and does not necessarily reflect the views of the World Bank group.

management and distribution, and eliminate the "free ride" to workers whose incomes were not closely linked to quality and quantity of labor. In its most extreme form, the complete household responsibility system (*HRS*), this aim has undoubtedly been achieved. Under the *HRS*, all cultivation activities for assigned pieces of land are the responsibility of a particular household. Where the team or other collective body supplies services, fees may be charged to the household. Households are still responsible for meeting their share of quota and above-quota procurement obligations, the agricultural tax, feed for collectively owned animals, and contributions to funds for collective welfare and accumulation, if any.

Despite the decentralized management, freedom to modify cropping programs is supposed to be limited even under *HRS*. For example, cotton and oilseeds acreage are specified by higher authorities, down to the team and household level. Grain quotas by variety are determined down to the team and household level. Timing and nature of operations such as plowing, sowing, and irrigation, as well as the seed varieties, are also supposed to be collectively determined, to permit efficient use of collective machinery and irrigation facilities. Thus in theory the decision-making responsibilities which have devolved are strictly limited, although practice sometimes exceeds what is allowed by theory.

Any radical institutional change is bound to introduce new problems. One concerns the management of team assets other than land, notably machinery. Especially under *HRS*, some machinery has gone unutilized, including that which was previously unworkable or too large in scale, or because households wish to save on operating costs; simultaneously, households are seeking to obtain their own draft animals to insure timely cultivation of the fields they manage. Other machinery has been divided up among groups or households, which at least leads to better maintenance. But when the team rents its machinery to work groups or families, there is no guarantee that it will be maintained and returned in good condition; it has been

suggested that a responsibility system should be devised for machinery management as well as field management. In the interim, demand for large-scale farm machinery has fallen sharply (along with the percentage of area tractor-plowed), while a market for custom services has arisen.

In the areas where implementation of the *RS* has not disrupted team management, it has presented few administrative problems, but in *HRS* areas the managerial situation has changed so drastically in such a short time that the administrative apparatus has had difficulty catching up. There is a danger that most government resources and services will be concentrated on the minority of farmers who choose to farm collectively, as in the early days of the producer's cooperatives. As local cadres come to accept the permanence of the *HRS*, it is to be hoped that the social infrastructure supporting agriculture will be appropriately reformed.

## II. Price Reform

The Chinese government in 1979 made major adjustments in the level and structure of farm prices, because of concern that the existing price structure provided insufficient income and also affected the choice of crops in undesired ways. Because these reforms forced sharply higher budgetary costs for consumer subsidies, the government has made it clear that no further major increases in the level of crop prices would be considered in the near future, although input prices could conceivably be adjusted downward if the profitability of the producing industries increases.

The current farm price structure encompasses four distinct prices: quota, above-quota, negotiated, and free-market prices. Quota prices apply to crops sold in fulfillment of procurement quotas; above-quota prices to crops sold in excess of the quota. The 1979 price reforms increased quota prices by 20.9 percent for grain, 15.2 percent for cotton, 25 percent for oilseeds, and 24.6 percent for pigs. The ratios of above-quota prices to quota procurement prices, which were 30 percent for grain prior to 1979, were in-

creased to 50 percent for grains and oilseeds, 20 percent for bast fibres, and 30 percent for cotton which previously received no above-quota premium (quotas for cotton were set at 1976–78 procurement levels). Above-quota prices do not apply to soybeans, the procurement price of which was increased 15 percent in 1979 and 50 percent again in 1981. Norms or obligations for above-quota sales are determined one or two months before each harvest, based on projected yields. Negotiated prices fluctuate at approximate parity with free-market prices. They apply to sales by individual farmers or (rarely) teams to the state which could otherwise legally be sold on the free market. In 1980, negotiated prices seem to have been 10–20 percent higher than above-quota prices. Overall, in 1979 the average procurement prices (weighted average of all four types of prices) increased by 25.5 percent for grain, 40.8 percent for oilseeds, 17.6 percent for cotton, and 37.0 percent for pigs.

Procurement quotas for most products were last fixed in the early 1970's, technically for a period of five years, but in fact they have not been adjusted (except downward in some areas in 1979). They are fixed for two major product categories—grain and cotton. Quotas for other crops, such as oilseeds, vary from year to year. If no quota exists for a product, the above-quota price normally applies. Due to these rules, increases in procurement are generally at higher than average prices and impart an uptrend to the weighted average price even with an unchanged price structure. The effect can be sizeable; of total value of net procurement of products of agriculture and sideline industry in 1981, 58 percent was at quota prices, 21 percent at above-quota prices, 12 percent at negotiated prices, and 9 percent free-market sales to the nonfarm population; the increase over 1980 was only 2.5 percent for quota procurements, but 30–38 percent for each of the other three components. As a result, the payment for each incremental ton of grain procured was 16 percent greater than the average price (66 percent greater than the 1979 quota price); for oilseeds, 18 percent greater than the average; except for cotton,

which was 8 percent less than the average (13 percent greater than the 1979 quota price), due perhaps to north-south price differences.

Does the new price level provide farmers with an adequate livelihood? The price level here is defined as the relationship between average prices of farm products and prices of industrial products. There is no objective standard of adequacy for the price level, inasmuch as it determines the distribution of income between farm and industrial sectors and perhaps between consumption and investment. Certainly it is true that procurement prices are lower than the level required to induce voluntary sales of an equivalent amount of produce. That is, they do involve a tax on the farm sector. However, the procurement quota, since it is fixed in absolute amount but varies among teams in proportion to land productivity and cultivated land per capita, satisfies the criteria which economists have set for the ideal land tax. Eliminating the quota procurement system would reduce revenues available to the government, either directly through increased procurement costs or indirectly, as an increased urban cost of living forced compensating wage increases, leading to decreased industrial profit. Therefore one cannot appraise the farm price level without considering the entirety of Chinese macroeconomic policy. For example, it is arguable that investment of industrial profits in order to augment the supply of consumer goods and expand employment opportunities in rural industry and commerce would increase real farm incomes and improve rural morale at less cost to the nation as a whole than a further increase in farm prices.

Does the new farm price structure provide an appropriate set of incentives? In reviewing price structure, the marginal price received or paid deserves our attention, because cropping patterns and input use are determined by comparisons of marginal revenues and marginal costs. Because most of China's agricultural products and modern inputs are internationally traded goods, the price structure is best compared to the structure of border prices (import or export prices adjusted for transport, trade and processing

TABLE 1—PRICE RELATIVES OF CROPS AND FARM INPUTS, CHINA 1980 (YUAN/TON)

1980 Price	Wheat	Maize	Rice	Soy	Rape	Cotton	Urea
Trade price <sup>a</sup>	\$ 191	125	373	296	326	2,070	222
Border price <sup>b</sup>	Y 406	294	372	556	607	3,850	477
Marginal domestic price <sup>c</sup>	500	360	408	690	1,020	4,500	552
<i>Wheat as numeraire:</i>							
Border price	1.00	0.72	0.92	1.37	1.50	9.48	1.17
Marginal domestic price	1.00	0.72	0.82	1.38	2.04	9.00	1.10
<i>Urea as numeraire:</i>							
Border price	0.85	0.62	0.70	1.17	1.27	8.07	1.00
Marginal domestic price	0.90	0.65	0.74	1.23	1.84	8.14	1.00

<sup>a</sup>1980 import prices FOB at ports of origin, except rice which is estimated 1980 export price FOB China.

<sup>b</sup>Trade prices, converted at Y1.7 = US\$1 and adjusted for transport, processing costs, and milling losses to the equivalent of a farm gate price.

<sup>c</sup>Above-quota price for crops except soybean; retail price to the farmer for fertilizers. Domestic price for urea is average of cost of nutrient equivalent from local products.

margins to the farm gate level). In most cases, if the relative marginal prices of crops and the ratios of domestic product prices to input prices differ from the corresponding relative border prices, a government could increase the total product supply from domestic and international sources without increasing its procurement costs. For example, as long as the domestic rice/fertilizer price ratio is lower than the border price ratio, it should be possible to increase the domestic rice price, exchanging part of the induced additional rice production for imported fertilizer, and increase national income in the process.

Since quotas are based on levels of local production which have been regularly attained in the past, crop producers can normally expect that any increments in production will be either sold at above-quota prices or better, or else (in part) retained, in which case the product may be sold at negotiated prices or on the free market. Hence marginal prices are not quota procurement prices, but some average of the other three prices. On this basis, Table 1 demonstrates that the 1979 price reforms brought about a price structure which should signal approximately the same cropping and input use decisions

for major traded crops as would occur if producers received border prices. With the price of nitrogenous fertilizer as numeraire, the price relatives for wheat, rice, maize, soy, and cotton based on border prices are 0.85, 0.78, 0.62, 1.17, and 8.07, respectively. Taking above-quota prices as the closest approximation to a marginal domestic price, the corresponding relatives are very similar: 0.90, 0.74, 0.65, 1.25, and 8.15, respectively. The relative prices of oilseeds such as rape and peanuts, which are high compared to border prices, are an exception and this may account for the rapid growth in area planted in these crops.

However, not all production teams face such near-optimal price signals: domestically produced nitrogenous fertilizer varies widely in price per unit nutrient, and teams depending heavily on high-priced fertilizers such as ammonium sulphate have an incentive to use less than the optimal amounts. The same applies to teams which normally cannot meet their procurement quotas, because of increased population or deterioration of crop yields, and therefore face quota procurement prices at the margin. More commonly, teams which farm with excessively high levels of input use (and there-

fore which have input marginal productivities lower than those in other countries) may find that marginal costs exceed marginal revenues, even though their marginal sales are at above-quota prices. These three cases suggest three different solutions: equalize fertilizer prices based on nutrient-equivalence, reduce unrealistic quotas, and reduce input use where the economic optimum is exceeded.

### III. Productivity Effects

It is now more than two years since the major price reform, and the new *RS* organizational forms have begun to stabilize. We can now assess the short-run effects of the implementation of these reforms through a comparison of 1978–1981 statistics, although it is not possible to distinguish their effects from those of administrative measures taken simultaneously.

The most remarkable effect has been the rapid expansion of acreage in cash crops from 9.6 percent of sown acreage in 1978 to 12.1 percent in 1981, at the expense of some decline in grain area (80.3 to 79.2 percent) and a larger decline in green manure area (6.1 to 4.6 percent). During the same period, increases in crop yields, which were larger for economic crops than for grains, have been observed across the board for every major crop. Only wheat and tuber yields fell slightly short of records in 1981.

These gains have been associated with the resumption in certain localities of old patterns of specialization in crops such as cotton, peanuts, or sugar, although increased area accounts for only part of production growth. For example, the doubling or tripling of cotton production in 1979–80 alone in the old northern cotton producing areas of Shandong, Hebei, and Henan is remarkable. However, the 1978–81 increases in per hectare yields in these provinces, at +134, +100, and +54 percent, respectively, outweighed changes in crop area of +73, –5, and +15 percent. The distribution of a high-yielding cotton variety may take credit for much of the yield increase.

In those provinces where the *HRS* or major changes in cropping pattern have been implemented, there has been a marked broad productivity effect. Nationally the portion of the gross value of agricultural output from crop cultivation increased only 4 percent in 1979–81, but in Shandong, Siquan, Honan, Anhui, and Guangdong the increases were 15, 17, 19, 30, and 54 percent, respectively (in constant prices).

Changes in procurements are difficult to interpret because a “breathing spell” was intended in which farmers would retain a larger proportion of their produce; moreover, the majority of procurements are planned and semicompulsory. Nevertheless the fully voluntary proportion (those purchased at negotiated prices or sold on the free market) has grown substantially. If one regards increases in procurements as a supply response to increased prices, then procurements of grain, cotton, oilseeds, and pork increased (1978–81) by 35, 37, 146, and 42 percent, and the corresponding elasticities with respect to average farm prices were about 0.8, 1.0, 2.4 and 1.0, respectively.

Finally, the official surveys of sources of income and consumer expenditure, as well as personal observation, suggest that due to the responsibility system, the number of man-hours expended in cultivation was reduced without comparable reduction in work accomplished, and this labor was diverted to side activities, such as house building or commerce. Peasant parsimony insured some reduction in costs of production, such as through the abandonment of inessential or costly machinery.

Thus a cursory look at recently released Chinese statistics suggests that the 1979–81 reforms can take credit for substantial increases in crop output in some provinces and a lesser increase nationwide, together with impressive growth in crop procurement. Their impact on side activities and peasant incomes has been even more obvious. However, one-time price and organizational reforms are not a reliable source of continuous productivity growth. In the future, the government must look to its third instru-

# Economic Reforms and External Imbalance in China, 1978–81

By BRUCE L. REYNOLDS\*

This paper reviews the relationship between foreign trade and China's domestic economy during the period 1978 to 1981, from a macroeconomic perspective. I first present a simplified version of a model constructed by Richard Portes (1979), and then use it to consider the relationship between internal and external balance in China. The model clarifies the tradeoffs which Chinese policymakers faced in this period, and the adjustment mechanism which brought a sharp disequilibrium in 1979 under control by 1982. In a third section, I review the sharp changes in development strategy and in institutional arrangements ("readjustment" and "reform") which occurred in late 1978 and 1979. This leads to a somewhat different view, compared with the implications of the Portes model, of China's prospects for continued growth through technology acquisition.

## I. The Portes Model

The Portes economy produces a single output, which may be used for consumption ( $c$ ), government purchases ( $g$ ), or exports ( $e$ ). While  $e$  and  $g$  are determined by the state, demand for consumer goods ( $c^d$ ) is a function of the real wage ( $w$ ) and money balances ( $m$ ), and may be greater or less than the quantity of consumer goods which the state places on the market.

Labor supply is also a function of the wage and money balances. Labor combines with a fixed capital stock ( $k$ ) to produce value-added ( $y$ ), which then combines with imports ( $i$ ) to produce output ( $q$ ). All imports are intermediate goods, and the level of

imports is determined by the level of domestic output.

$$(1) \quad 1^s = 1^s(w, m),$$

$$(2) \quad y = y(1^s, k),$$

$$(3) \quad q = \min(y, ai).$$

The planners seek to maximize a function  $U(c, g)$ . They have three policy instruments ( $e$ ,  $g$ , and  $w$ ) but face two constraints: that the consumer goods market be balanced, and that exports be sufficient to meet a trade balance target  $\bar{B}$ . In other words, once one policy variable is specified (say,  $g$ ),  $w$  and  $e$  are determined by the need for internal and external balance.

The planners' situation may be illustrated graphically by collapsing this model into two equations in  $e$  and  $w$ , showing the locus of points which satisfy the requirements for internal and external balance, respectively:

$$(4) \quad CC: e$$

$$= y(1^s(w)) - c^d(w) - g, \quad de/dw < 0;$$

$$(5) \quad BB: e$$

$$= \bar{B} + (1/\alpha)y(1^s(w)), \quad de/dw > 0.$$

Equation (4) states that if available consumer goods are to equal  $c^d$ , given  $w$ , then exports may not exceed the residual after consumer demand and government purchases are subtracted from output. Equation (5) has two terms. The first is the trade balance target. Even if imports are zero, exports must be sufficient to meet this target. The second term shows that at any given wage  $w$ , labor supply will presumably be nonnegative, generating value-added  $y$ ; multiplying by  $1/\alpha$

\*Associate professor of economics, Union College.

TABLE 1—ANNUAL PERCENTAGE CHANGE IN KEY INDICATORS

	1978	1979	1980	1981
(1) $g/GNP$	7.6	4.4	-1.4	-2.9
(2) $e/GNP$	1.2	1.5	1.6	1.3
(3) $(w/p)$	5.2	6.5	6.5	
(4) $P$	0.7	1.9	7.5	
(5) Trade Balance (\$B)	-1.7	-2.0	-1.2	0.0
(6) $GVAO$	9.0	8.5	2.7	3.3
(7) $GVIO$	12.2	8.5	8.7	4.0
(8) Light Industry	18.6 <sup>a</sup>	9.6	18.4	13.6

Notes: (1) Annual increase in government expenditure (at all levels), divided by previous year *GNP*, in current *yuan*; (2) Annual increase in exports, divided by base year *GNP*, in current dollars; (3) Annual percentage increase in average wages to all industrial workers and staff, deflated by official price index for consumer goods; (4) Annual percentage increase in official price index for consumer goods; (5)–(8) Annual percentage increases in gross value of agricultural output, industrial output, and light industrial output. All figures based on official Chinese statistics, except as noted.

<sup>a</sup>Based on author's estimate of 1977 light industrial output.

gives the quantity of imports needed as intermediate goods to accompany this value-added in producing output. Exports, then, must increase by this amount to achieve external balance.

Under plausible assumptions, an increase in  $w$  is associated with a decrease in  $e$  along the *CC* curve and an increase in  $e$  along the *BB* curve.

## II. Interpreting Chinese Macroeconomic Experience

Let us consider recent Chinese macroeconomic experience in the light of this model. In 1977, the trade balance was slightly positive, and there was no evidence of strong inflationary pressure on the domestic market for consumer goods. Let us suppose, then, that in 1977, economy was in equilibrium, with exports at  $e_0$  and the real wage at  $w_0$ .

In 1978, as the new Chinese government moved sharply away from political struggle and toward Deng Xiaoping's economic agenda, government spending increased by a remarkable 7.6 percent of *GNP*. (See Table 1.) Such a change shifts *CC* down, leaving *BB* unchanged. To maintain internal and external balance would have required an appropriate decrease in  $e$  and  $w$ , to some new "planners' equilibrium" ( $e_1, w_1$ ). Instead, both  $e$  and  $w$  rose sharply. Chinese policy,

according to the Portes model, was moving away from balance, creating the possibility of unbalanced trade and the certainty of excess demand on the domestic market. In 1979, this pattern was repeated: another marked increase in  $g$ , accompanied by rising  $e$  and  $w$ .

This two-year expansionary burst precipitated exactly the difficulties which the model would predict. The trade balance deteriorated rapidly, showing a \$3.7 billion deficit for the two years taken together. Imbalance on the domestic side, as measured by the official price index for consumer goods, appeared more slowly; in 1978, the 0.7 percent inflation rate was not much above the historical trend for the *PRC*. But by 1980, very strong upward pressure on prices is clearly reflected in Table 1.

To reestablish external and internal balance, the Chinese government shifted to a contractionary macroeconomic policy. Government spending decreased in 1980 and again in 1981. Although wage payments continued to rise in 1980, increases in 1981 were more modest (and in both years, actual inflation may have eaten up more of the money increase than official statistics indicate). The year 1982 saw a continuation of this policy: a planned decrease in  $g$  comparable to that in 1981, and a new policy tying wage increases to productivity increases. As the

model would predict, the trade deficit disappeared in the wake of these policies, and inflationary pressures subsided considerably.

In terms of the Portes model, the economy began in 1979 at a point above the *CC* curve and below the *BB* curve, that is, at a point of both external and internal imbalance. By lowering *g*, the *CC* curve was shifted upwards to restore the internal balance. However, this left the external imbalance with exports falling short of imports. One alternative would have been to raise exports by this amount (at the cost of further contracting *g*). Instead, the Chinese have evidently lowered their trade balance target, accepting a modest external debt of some \$4½ billion, thereby shifting *BB* downward toward *D*.

The Portes model, then, proves to be a useful paradigm through which to enhance our understanding of China's recent macroeconomic experience. It highlights the trade-offs facing Chinese policymakers, for example that between wage increases and external balance. It shows us a cycle of expansion-imbalance-contraction-adjustment which, though set in the context of a planned economy, is similar to its analogue in market economies. The Chinese response to imbalance in 1980 and 1981 appears prompt and forceful.

This treatment, however, leaves completely unaddressed the major policy shifts of late 1978 and late 1980 which are usually considered central to any understanding of this period: that is, the advent of "readjustment and reform," and the retreat from reform two years later. Let us review 1978–81 from this perspective, using a more eclectic and institutional approach, and consider whether important features are added to the story.

### III. Readjustment, Reform, and Technology Acquisition

China in 1977–78 faced problems strikingly similar to those of Poland in the early 1970's. (For an excellent review of the Polish case, see John Montias, 1982.) Like Poland, China needed to absorb a surge of new entrants to the labor force (4 to 5 million per year, in addition to millions of "rusticated youths" returning to urban areas). Like the

Gierek regime, Deng Xiaoping's legitimacy hinged on perceived economic successes, especially an increase in consumer goods. In addition, labor productivity in industry was stagnating. Like Poland, China's response to this set of problems was a massive program of technology imports. A (now-defunct) ten-year plan was proclaimed, featuring 120 major projects. Vice-premier Li Xiannian talked enthusiastically of \$600 billion in new capital requirements. Awash in credits from eager OECD lenders, China went on a shopping spree for complete plants and equipment.

The parallel with Poland extends to the failings of this program. The Chinese program favored the iron and steel sector, which was unlikely to generate the foreign exchange needed to balance trade. The investment projects were capital intensive, with long gestation periods and low potential for labor absorption. As in Poland, project approval often came through informal political processes at the highest levels of government, with insufficient consideration for the feasibility of the project or the need for infrastructural investments. This was especially true in the last two months of 1978. As criticism grew, it became increasingly likely that the Third Party Plenum, scheduled for December, would reverse the policy. In an astonishing breakdown of any semblance of investment planning, twenty-two foreign equipment contracts, costing US\$2 billion, all of it outside the state plan, were signed in the three weeks prior to the Plenum, often on the basis of marginal notes and encircled words on reports by a few top officials.

The December 1978 Plenum, and the June 1979 People's Congress, introduced a policy which reacted to this situation in two ways: readjustment and reform. Both constituted departures from the traditional Stalinist economic model. Readjustment would modify the excessive stress on heavy industry, shifting away from the Stalinist reliance on high investment rates and extensive growth. Reform would modify the excessive centralization of economic power, giving enterprises and local (i.e., provincial and municipal) governments more authority over production and investment decisions.

What were the implications of "readjustment" for internal and external balance? Under the new policy, the investment rate would be brought down from above 35 percent to between 25 and 30 percent. China would increase the share of consumption in *GNP*, but would have to accept a reduction of the industrial growth rate from above 10 percent to perhaps 6 percent. If one were to model this discontinuous shift from one long-term growth path toward another, one would find that the producer goods sector necessarily approaches its new (lower) long-term growth rate from below, while the (manufactured) consumer goods sector approaches it from above. Intuitively, the capital stock needed to generate 6 percent industrial growth is smaller relative to *GNP* than that required for 10 percent growth. Thus China's heavy industry sector (approximately the same as producer goods) could stand still for several years, waiting for *GNP* to catch up; light industry (chiefly consumer goods), meanwhile, could use at least some of the released resources to spurt ahead. As Table 1 shows, the relative growth rates of heavy and light industry during the past four years have in fact followed this path. In terms of the Portes model, the Chinese economy, as it transits from its 1978 to its "readjusted" growth path, is able to provide a one-time dividend of wage goods which shifts *CC* upward to the right, greatly easing the constraints on policymakers. (Of course, the other side of this coin is that the dividend will peter out when readjustment is complete.)

Meanwhile, economic reform was shifting some powers to enterprises and localities. Enterprises began to pay bonuses out of retained profits, with a significant impact on the total wage bill (but not on incentives, see my 1982 study). Localities began to invest in locally controlled light industry, with an eye to producing the consumer goods which central policy (and the profit motive) dictated. Thus just as the readjustment policy began to curb excessive central government investment, reform generated an investment boom by local governments. This, along with decentralization of foreign trade powers, exacerbated the balance of trade deficit in 1979 and 1980, and generated strong inflationary

pressures. In late 1980, some Chinese economists spoke privately of their fear that the center had "lost control of the economy." Chen Yun, China's economic *eminence grise*, warned in December that "it is not yet clear that the Polish incident cannot happen in China."

Clearly, readjustment facilitated the search for macroeconomic balance, while reform greatly hindered it. Aware of this, the leadership reinforced the former and scrapped the latter, beginning in early 1981. Enterprise powers withered. Local government budgets were recentralized. (Their share of state investment funds has fallen from its 1981 level of 51.5 percent to only 22 percent of a lower 1982 budget.) Foreign trade has also been brought under strong central control. One result: imports in current dollars from OECD countries were down 17 percent in the first half of 1982.

While the triumph of readjustment over reform has solved the problem of internal and external balance, it leaves the economic system where it was in 1978. What then of Deng Xiaoping's program for modernization via technology acquisition? Pursuing it in 1977-78 in a centralized fashion led to irrational excesses by central ministries. Pursuing it in 1979-80 in a decentralized fashion led to irrational excesses by local governments. The sharply declining trade figures in 1982 suggest that the strategy has been placed on ice. In the light of China's experience from 1978 to 1981, one can understand why.

## REFERENCES

- Montias, John M., "Reviewing the Polish Experience," *Bulletin*, Association for Comparative Economic Systems, Fall-Winter 1982.
- Portes, Richard, "Internal and External Balance in a Centrally Planned Economy," *Journal of Comparative Economics*, December 1979, 3, 325-45.
- Reynolds, Bruce L., "Reform in Chinese Industrial Management: An Empirical Report," Joint Economic Committee, *China Under the Four Modernizations*, Washington: USGPO, 1982, 119-38.

# Enterprise-Level Reforms in Chinese State-Owned Industry

By WILLIAM BYRD\*

By the late 1970's, China had built up a comprehensive modern industrial system, predominantly state owned, which was able to produce a wide range of producer and consumer goods domestically. China's long-run industrial growth rate has been one of the highest in the world. This high growth rate was maintained even in the decade 1967-76, now vilified as a period in which "leftist" ideological considerations dominated economic policy. But Chinese industry has suffered from fundamental weaknesses, which have made it very inefficient. The most important and chronically debilitating problems have been: inefficiency in converting material inputs and energy into outputs; poor product quality; mismatches between supply and demand; excessive inventories; inefficient investment; widespread underutilization of fixed assets; and an overly dispersed, inefficient regional pattern of industrialization. There is evidence not only of failure to improve efficiency, but of actual deterioration in performance over time, at both macro and micro levels. By the late 1970's, problems of inefficiency had become so serious that they were an important factor in precipitating economic reforms.

This paper reports on enterprise-level reforms in Chinese state-owned industry, which were implemented starting in late 1978. The basic components of the reform program and the main trends that have emerged during reform implementation will first be discussed. Then the success of reforms in im-

proving efficiency will be evaluated. A number of conclusions arrived at will differ from the judgements of other observers. Reforms have rapidly expanded in scope, to the point where the bulk of the state-owned industrial sector has become involved. Reforms have had a considerable impact on the orientation and activities of Chinese managers. Reforms have been associated with a perceptible improvement in production efficiency during the past several years.

## I. Basic Components and Main Trends in Implementation

Implementation of enterprise-level reforms has often appeared chaotic and haphazard, more the result of *ad hoc* responses to immediate exigencies than of a comprehensive plan. Various pilot programs have been instituted for different kinds of enterprises and in different parts of the country; they differ significantly from each other in terms of some of their basic provisions. Nevertheless, reforms are related by their common goal of improving efficiency. Moreover, all reform programs have in common a number of basic components: 1) devolution of greater discretionary authority to enterprises in production and investment activities; 2) use of material incentives (in the form of profit retention schemes for enterprises and bonuses for individual workers) to supplement administrative directives in guiding enterprise decision making; and 3) an expanded role for the market mechanism in resource allocation. All three involve drastic departures from policies prevalent during the period of "late-vintage Maoism" (1967-76). Throughout that decade, material incentives at enterprise and individual levels in the state sector were strictly forbidden, the market played virtually no role in the allocation of producer goods and an extremely limited one in the allocation of consumer goods, and enterprise management had considerably less decision-making power.

\*Ph.D. candidate department of economics, Harvard University. Financial assistance from a Schumpeter Prize Fellowship and Foreign Language Area Studies awards is gratefully acknowledged. Comments were received from R. Dernberger, J. M. Montias, Dwight Perkins, and Barry Naughton. Responsibility for views expressed is my own and should not be attributed to any other person or institution. The paper uses information from Chinese journal articles and statistical, economic, and legal compendia. Non-Chinese scholars who have written on this subject include Robert Dernberger, Barry Naughton, and Bruce Reynolds.

The most striking feature of reform implementation has been rapid, often uncontrolled growth of participation in the various reform programs. Reforms started in late 1978 with six pilot firms in Sichuan Province. The number of enterprises involved increased rapidly, reaching 6,600 by mid-1980. Though comprising only about 11 percent of the total number of state industrial enterprises, they accounted for 44 percent of the gross output value and 57 percent of the profits of this group. The campaign to promote "economic responsibility systems" begun in April 1981 brought on a new wave of expansion. By the end of the year, 80 percent of all state-owned industrial enterprises were said to be involved. This hurried implementation undoubtedly has resulted in simplifications and distortions. Nevertheless, the Chinese accomplishment in engineering such a rapid transformation of the state industrial sector is remarkable.

A second outstanding feature has been significant local and provincial control over the reform implementation process. Sichuan Province was a pioneer in this regard, but other territorial units also exercised considerable initiative in the early stages. Later, in the campaign to promote economic responsibility systems, decisions on key provisions affecting enterprises were formally devolved to their immediate superiors, the industrial bureaus and corporations. These could decide both enterprises' quotas for profit remittances to higher levels and their retention rates for above-quota profits.

Another salient trend has been the increasing focus on industrial bureaus and corporations as opposed to enterprises. To some extent this may simply reflect retrogression—elements of the bureaucracy which played a dominant role in the prereform situation have attempted to regain control. But there are sound economic reasons why in certain spheres decentralization should not proceed all the way down to the enterprise level. This is particularly true of investment decisions and control over investment funds, in the absence of effective financial intermediation by the banking system. Centralization at the bureau or corporation level can avoid duplicative, inefficient investments and resulting excess capacity, at least within the

area of the bureau or corporation's jurisdiction. In production as well, a certain amount of centralization can eliminate the most obvious forms of local duplication and waste. On the other hand, centralization of financial resources and rewards undoubtedly harms incentives at the enterprise level. Bureau-level reform programs generally try to maximize the benefits in this tradeoff by maintaining control over investment funds, while allowing funds for workers' bonuses and other benefits to be allocated directly by the enterprises.

Turning to the new financial incentive schemes, five main trends are apparent: 1) progressive simplification; 2) an increasing focus on profits as the only important performance indicator and reward determinant; 3) a shift in the basis of evaluation from comprehensive plan fulfillment to growth (of profits) to arbitrary quotas for profit remittances to supervisory authorities, determined at relatively low levels in the economic management system; 4) over time, progressive increases in enterprises' marginal profit retention rates; and 5) growing but largely ineffective concern with stabilizing government revenue from enterprise profit remittances.

Without trend 1), the new financial incentive systems could not have been so universally applied so quickly. The trend toward near-exclusive focus on profits and neglect of other targets may allow enterprises to manipulate product mix, to the detriment of central goals. On the other hand, growth of profits (not the absolute level) may well be the best single target indicator to use. Trend 3) is a reflection of China's relatively weak planning system and the increasing importance of lower levels. Though use of arbitrary quotas may result in more flexibility, the scope for bargaining is widened immensely. The trend of increasing profit retention rates has caused enterprise financial resources to grow rapidly, while government revenue from enterprise profit remittances has fallen.

## II. Impact on Efficiency

Despite the problems noted in the previous section, reform implementation has been

rapid and, under the circumstances, surprisingly thorough. In the area of material incentives, reforms appear to have been successful in changing the orientation of enterprise managers and other decision makers from increasing physical output to increasing profits. The expanded role of markets and market-like activities in resource allocation also is well documented. Progress has been slowest in the area of devolution of discretionary authority to enterprises. But two important changes have occurred: the increased use of economic as opposed to other criteria in decision making at all levels; and the shift in locus of authority over day-to-day operations from the Party Committees of units to their directors.

Reforms can thus be judged successful in narrow terms: they have had a significant impact and have contributed to a major change in orientation. Have they also been successful in improving efficiency? This question cannot be answered definitively because of our inability to separate the effect of reforms from that of other factors and policy changes. Nevertheless, it will be shown that reforms have been associated with significant improvements in the production efficiency of Chinese state industry. In 1981 virtually the entire sector came to be affected by reforms. A large majority of enterprises participated directly; even those that did not are eligible to draw some discretionary financial resources based on their performance, and most firms can freely market their above-plan output. Therefore changes in the efficiency of the state-owned industrial sector as a whole do provide relevant evidence for an evaluation of the success of reforms.

It remains to be determined which indicator of efficiency is most appropriate in the Chinese context. A direct measure of the efficiency with which inputs are converted into outputs is the ratio of current input consumption to gross value of output. *Ceteris paribus*, changes in this ratio are indicative of changes in production efficiency. The problem with this indicator is that it is subject to bias from a number of sources, the most important of which are: price changes for goods purchased from or sold to units outside of the sector; changes in the relative proportions of goods with different ratios of

input consumption to output value produced; and changes in the extent of vertical integration. Prices of industrial goods produced by the state sector on average have not changed greatly, rising by 0.7 percent between 1970 and 1980. Adjustments have to be made, however, to take into account the sharp 38.5 percent rise in agricultural procurement prices in 1978–81, which must have had a considerable effect on industrial costs. Changes in industrial structure would cause reductions in the ratio of input consumption to gross output value to overestimate actual improvements in efficiency if production became more vertically integrated or if methods of reporting changed, with some inter-enterprise transactions netted out. It is extremely doubtful whether either of these eventualities has actually occurred. Changes in the composition of industrial output appear not to have caused any serious bias, since in 1978–81 production of goods with a high ratio of input consumption to gross output value on average grew slightly faster than that of goods with a low ratio.

There was a perceptible decrease in the estimated ratio of current input consumption to gross output value of Chinese state industry between 1975 and 1978, and in each of the three years 1979, 1980, and 1981. The ratio fell from 0.650 in 1978 to 0.632 in 1981, a reduction of almost 3 percent. If the increase in agricultural procurement prices is taken into account, the improvement in efficiency becomes more striking. It is reasonable to assume that the average increase in prices paid by the state-owned industrial sector for inputs from agriculture was in the range of 10–20 percent. Some agricultural price increases were absorbed by the commercial system and not passed on to industrial producers, while others were offset by corresponding output price increases, and still others may have been compensated for with subsidies. Depending on the estimate of the total value of agricultural products consumed as material inputs by state industry, the reduction in the ratio of “real” input consumption to gross output value was between 4 and 7 percent. The assumptions used in arriving at these estimates are highly conservative, so they probably understate the actual improvement in production efficiency.

Changes in other indicators of efficiency have been less clearcut. Labor productivity has stagnated, but since Chinese enterprise managers still have very little control over the size of their work forces, this does not reflect the impact of reforms. Inventories have not fallen dramatically. There has, however, been an important shift in the composition of inventories: stocks of inputs have fallen or grown only slightly, while inventories of final goods have risen sharply. This reflects a dramatic change in market conditions from the predominantly seller's market prevailing in the past to a buyer's market for many manufactured goods.

Capital productivity in the state-owned industrial sector has declined. The ratio of gross industrial output value to the total value of fixed assets dropped from 1.03 in 1978 to 1.01 in 1980 and 0.96 in 1981. This phenomenon is at least in part the result of difficulties reforms have encountered in attempting to improve the efficiency of investment. There are two main sources of problems: capital is still relatively cheap—interest rates and capital charges are low, and their impact is even smaller in cases where profit retention rates are low; and overinvestment has been encouraged by the growth of enterprise discretionary funds and increased availability of bank loans.

Capital construction investment of all types financed by enterprises and local governments has grown rapidly, from 17 percent of the total in 1978 to 30 percent in 1980 and 33 percent in 1981. Investment financed by domestic and foreign bank loans also has risen substantially since 1978 and in 1981 accounted for nearly 19 percent of the total. The share financed by central government budget appropriations dropped from over 80 percent in 1978 to less than 50 percent in 1981. These figures understate changes in the source of financing of fixed investment in state industry. They do not include investment in renovation and modernization of fixed assets, financed mainly from local government and enterprise resources and bank loans, which has grown rapidly in recent years.

This decentralization of investment financing has had serious adverse consequences for

the structure of China's investment program. Much investment financed by enterprises and localities has been duplicative and wasteful. In 1981 total investment in capital construction was cut back sharply in the interest of macroeconomic stability. To do this, the central government was forced to reduce disproportionately those parts of the investment program over which it had the greatest control—large projects undertaken by central organizations. As a result, cutbacks were especially severe in key infrastructure sectors like energy and transportation, which suffered reductions of nearly 21 percent and over 35 percent, respectively. These most likely have had a detrimental effect on the future growth potential of the Chinese economy. Though some projects suspended in 1981 have since been reinstated, the cost in time lost and resources idled has probably been high. The adverse impact has been exacerbated by a shift in the composition of capital construction investment from projects that increase future production potential to so-called "nonproductive" investment (mainly housing), which rose from 17 percent of the total in 1978 to over 41 percent in 1981.

In China since 1978 industrial production by the state sector has become more efficient in the sense that fewer inputs are required to produce the same amount of output. Productivity of factors, however, has stagnated, and at the macro level the overall pattern of China's investment program remains highly inefficient. Neither of these contradictory trends can be attributed entirely or even primarily to the impact of reforms, but reforms have played a significant role in both, and deserve some of the credit and blame. Introduction of material incentives at individual and enterprise levels has helped improve production efficiency. The resulting increase in enterprise discretionary financial resources, however, has stimulated investment demand and led to severe problems in controlling both the level and composition of investment. To achieve further progress, a way must be found to reap the benefits of the new incentive systems while at the same time avoiding the adverse impact on investment.

## INTERNATIONAL TRADE

### 'De-Skilling,' Skilled Commodities, and the NICs' Emerging Competitive Advantage

By ALICE H. AMSDEN\*

The dynamics of comparative advantage are such that the level of productivity in a sector in country *A* may be lower than in country *B*, but *A* may out-compete *B* if *A*'s rate of growth of productivity is higher. In what follows, I bet on the skilled branches in the newly industrializing countries (*NICs*) as the dark horse of international trade in the belief that gains in productivity are greatest in such fields. I show why this might be so in terms of a paradox: how it is that the transfer of technology can be most effective in a sector where the process of "de-skilling" has progressed the least. Because de-skilling has been far from complete, a world of three inputs is reality and trade among developing countries (South-South trade) is misperceived if three inputs are collapsed into two.

The high skill content of South-South trade, corroborated in recent empirical work, is taken as presumptive evidence of the *NICs* emerging competitiveness in the skilled manufacturing branches. My interpretation of the small body of literature on the large question of technology in the *NICs* is designed to illuminate why such countries compete better in fields where the human factor is greatest, compared to other industries also of recent origin in such countries which are machine paced or process centered. The *NICs* competitiveness in industries which rely upon large quantities of unskilled labor has long been demonstrated in exports to the North, so they are ignored here.

#### I. The Tendency to De-Skill, but the Persistence of Skilled Workers and Skilled Sectors

A lively debate has been sparked by the work of H. Braverman (1974) on the issue of

whether or not there occurs a process of de-skilling as countries accumulate capital and advance technologically. The debate is of interest to trade theorists insofar as it bears upon technology transfer, and hence the commodity composition of trade. Although no one appears to have considered explicitly the relationship between de-skilling in the industrialized economies and the transmission of production methods to newly industrializing ones, it seems reasonable to imagine that where de-skilling occurs, there emerges a two-factor world (labor and capital), with skills as a third element in production dropping out of the picture. The transfer of technology might be said to be facilitated with de-skilling insofar as one might suppose the skilled trades to be least amenable to international diffusion: knowledge of the production process stays locked in the experience of skilled workers, despite the efforts of management to wrest control from them (Braverman, 1974), and skilled production processes, more than most others, remain "tacit" rather than codifiable in blueprints (Richard Nelson and Sidney Winter, 1977).

The proposition of de-skilling follows directly from the writings of Smith and Marx on the division of labor. With the finer subdivision of tasks, the floodgates are opened to the application of machinery because specialized labor becomes simplified.

In certain respects, both the product cycle and Heckscher-Ohlin-Samuelson (H-O-S) models of international trade implicitly accept the de-skilling hypothesis. In the early stages of the product cycle, skills resurface temporarily with the advent of new technology, and there is a halt to technology transfer. In the end, however, the production process becomes standardized, skill content is

\*Department of economics, Barnard College.

reduced, we are back to two-factors, and technology transfer recommences. The two-factor implication of de-skilling gives the H-O-S model its explanatory power. Trade proceeds on the basis of a polarization of countries and commodities according to their relative endowments and intensities of capital and labor. When skills are considered explicitly, they are lumped together with capital (not labor), so that a two-way classification continues.

While technology transfer may accelerate with a decrease in manual skills, it may be arrested by an increase in intellectual skills, although it is not intuitively obvious where to draw the dividing line between the two. The problems of technology transfer once the complication of intellectual skills is introduced are either ignored (technology is assumed to be universal in the H-O-S model); or they are implicitly assumed to exist only when countries misbehave and skip over stages in the chain of comparative advantage (otherwise one stage of production automatically prepares the way technically for the next stage; Bela Balassa, 1977); or they are minimized, that is, the end of the product cycle is seen to involve relatively invariant technology which may be acquired either through foreign licenses or foreign investments.

While there may be a tendency towards de-skilling, reality is not quite so simple because science has not been wholly successful in transforming production. If one takes de-skilling and the "capital" in capitalism seriously, then one would expect the technological norm to be characterized by large amounts of capital per unit of unskilled labor. Yet this particular factor configuration appears to be the exception rather than the general rule. Certain commodities may be said to have altogether evaded capitalist production *techniques* (but not "relations") insofar as they are produced with little capital but with large doses of unskilled labor (for example, clothing and wood products). The techniques of other commodities (say, chemicals and basic metals) still require large quantities of both capital and skills. Finally, some commodities are typically produced

with little capital but with large amounts of skilled labor (nonelectrical machinery, ships).

Clearly, then, the realities of production indicate not the two-factor world which de-skilling would suggest but rather a three-factor one. As R. M. Stern (1976) and others have argued, the empirical evidence supports treating capital and skills not as substitutes but as complements, although not perfect ones.

While it may be warranted to aggregate capital and skills for certain purposes, for others doing so may prove distortive. This appears to be the case when one considers South-South trade in manufactures, which in 1979 accounted for 36 percent of manufactured exports from the South to the world.<sup>1</sup> Some have held high hopes for South-South trade insofar as they have viewed it as a way for southern countries to develop "intermediate" technology; to establish economic ties on which closer political bonds could coalesce; to raise productivity in more sophisticated manufactures; and to avert involvement in the depressions which recurrently grip the industrial heartlands. Others, however, have viewed it with a jaundiced eye: a two-factor generalized H-O-S model leads to the prediction of greater capital intensity (and hence resource misallocation) in manufactured exports to the developing than developed countries. In fact, it was argued early on that South-South trade in manufactures is marked by a heavy reliance on skilled labor (see my 1980 article). This has been corroborated by strenuous econometric exercises at the World Bank which include factor content analyses as well as regressions for more countries, commodities, years, and proxies for capital and skills. The results of either method show more unskilled labor per unit of exports to the North than the South. But the southern flow employs only marginally more capital and overwhelmingly more skills than the northern flow.

<sup>1</sup>In discussions of South-South trade, manufactures are variously defined. The 36 percent is based on manufactures defined as SITC 5-8 minus 68.

## II. Technological Creativity in the Skilled Sectors by Comparison with Sectors using More Capital

South-South trade may be taken as presumptive evidence of the emerging competitiveness of skilled industries. Thus, skeptics of "revealed comparative advantage" notwithstanding, I assume that the rate of growth of productivity of skilled industries in the South is greater than the rate of growth of productivity of the remainder of industries in the South, as well as greater than the rate of growth of productivity of skilled industries in the North relative to other northern industries. Given the unimportance to date of skilled exports from the South to the industrialized countries, it may be inferred that the fast rate of growth of productivity of such exports has only just begun (assuming that the markets of the industrialized countries are most difficult to penetrate).<sup>2</sup> There is even some evidence from an early period of a faster rate of growth of productivity in high skill sectors if such sectors are taken to be capital and durable consumer goods; that is, between 1968 and 1974, before the upsurge in capital goods production in many *NICs*, the rate of growth of productivity of capital and consumer goods in 16 *LDCs* was 3.3 percent compared to only 1.5 percent for mainly nondurable consumer goods and 2.6 percent for intermediate goods (UNIDO, 1979).

The assumption that the South is developing a competitive edge in skilled goods may strike an unrealistic note given the supposed problems of transferring the presumptively tacit technology of such goods. Further, the productivity gap between developed and developing countries has been predicted to be smallest the more process centered or machine paced the industry, and to be greatest the larger the role played by skills (W. Arthur Lewis, 1965). However: tests of the mechanization hypothesis have tended to be inconclusive (although conceivably because most

are based on weak data which predate the upsurge of manufacturing in *NICs* in the 1970's (Simon Teitel, 1981)); no tests look at rates of change of productivity although the dynamics of the situation warrant this; and none explicitly tests for the relationship between productivity differences and skills. Moreover, it is just as reasonable to imagine a priori that process-centered, machine-paced industries will be relatively least efficient as most efficient in the South. Because the division of labor is most highly developed in such industries, efficient production demands complex coordination at the level of the firm, complementary support services at the industry level, etc. For this reason, while productivity at the point of production may be expected to equalize internationally in process-centered and machine-paced industries, it may be expected to vary enormously at levels more remote from the pipeline or shopfloor.

To explain further why skilled goods may be gaining a competitive advantage, I focus on the technological characteristics of one component of South-South trade—capital goods. I do so despite the fact that many capital goods, even when defined to exclude passenger motor vehicles, are most notable for their high capital requirements (for example, heavy electrical equipment). These capital goods, however, tend in the general case to be traded far less than the others, especially nonelectrical machinery (C. T. Saunders, 1978). Capital goods, moreover, account for only 15–30 percent of total South-South trade in manufactures, depending on definitions. It will be assumed, nevertheless, that their production characteristics are generalizeable to other skilled exports.

The main finding of research on the technological activity of the *NICs*—for all industries and for indigenous firms mainly—is that it exists. The second finding is that firms resort at discrete turning points in their growth to foreign sources of technology. Third, a large part of the technological activity of the *NICs* is devoted merely to assimilating foreign sources. Another part is devoted to adapting foreign designs to suit

<sup>2</sup>It is possible that the *NICs* export a higher value of skilled manufactures to the North than the South, but this doesn't appear to be the case.

local conditions: descaling, converting from mass to batch production, reducing imported raw material requirements, etc. Fourth, adaptive engineering, as it might be termed, only rarely gives rise to innovations at the technological frontier. The modifications to which it does give rise, however, may account in some unknown part for the attractiveness of southern exports in southern markets. Further, small, incremental innovations are responsible for rather sizeable productivity gains over time. Fifth, the *NICs* have advanced much further in production engineering (efficient operation of plant) than in systems engineering (design of process and product). For the most part, there persists a dependence on foreigners for the execution of new projects.

I believe that the major reason why skilled branches do the best is that, paradoxically, the process of de-skilling has progressed the least. The proposition that technological tacitness and nonreplicability are greatest the more skilled the commodity appears to be true; but this is not without virtue. It means that the acquisition of technology from abroad in the metalworking branches is least likely to be accomplished through turnkey imports. (Turnkeys are popular the more standardized the product, for example, chemical plants, textile mills.) Instead, the lesser codification of capital goods technology forces firms to participate actively in the technology acquisition process—which is effected most frequently by copying or foreign licenses. Through a process of active participation in technology transfer, a more intimate knowledge of systems design is acquired. This allows for better adaptive engineering, out of which arises the small, anonymous improvements in productivity mentioned above. Design of process and product, moreover, is more intimately connected technically in capital goods industries than in others. Therefore, adaptive engineering is itself conducive to gaining a capability in design. Thus, there is an interconnection or “virtuous circle” between production engineering and project execution in the capital goods sector that is absent or less pronounced elsewhere: project execution with (perforce) local involvement leads to greater

adaptive engineering; and adaptive engineering and learning-by-doing lead to a greater ability to design. If the *NICs* are light years away from establishing a local capability in systems engineering in most of the newer import-substitution sectors, they are within range of doing their own designs in many capital goods branches; for example, ships; offshore oil rigs; boilers; *CNC* lathes; special purpose, agricultural, and textile machinery, etc.

The feedback effects to and from production engineering and project execution are feeble even in the capital goods sector when copying is the exclusive means by which technology is acquired. This is because the scientific principles underlying a design are unlikely to reveal themselves in the process of copying. This makes reverse engineering more difficult. Thus, firms which must rely on copying alone to obtain their technology may be expected to perform less well than larger, more financially able firms. In fact, contrary to expectations, a sizeable percentage of capital goods output in Brazil, India, South Korea, and Taiwan is realized in firms which are large even by international standards. The interconnection between capability in production and design, moreover, is not automatic: it is rare to find a capital goods builder in a *NIC* which evolved from small to large without recurrent injections of foreign know-how and continuous local investment in technological learning.

We know that the greater the division of labor, the greater the degree of de-skilling and the use of capital; and implicit in production systems which have advanced furthest along these dimensions is a greater application of science to both product and process design. In addition, with a highly elaborate division of labor, functions such as *R&D* and production become more specialized. Hence, the more capital using the good, the greater the separation of production engineering and project execution, and the less the possibility of proficiency in one leading to proficiency in the other. Ignorance of design, however, may prove as great a handicap to exporting as incapacity to serve poses to winning in tennis.

### III. Conclusions

It would be reductionist to attribute the impending competitiveness of skilled goods—which has already manifested itself in southern markets—to technological conditions alone. Differences in costs of skilled manpower and capital, in scale, and etc., may also distinguish the performance of sectors in the *NICs* which rely largely upon skills rather than capital. But, just as one cannot explain adequately differences in the trade patterns of the developed countries in terms of factor endowments, so, too, the existence and unevenness of the technological capability of the *NICs* must be acknowledged to explain their trade flows. This is all the more important given that capital in the *NICs* cannot be taken as homogeneous: the technological element intrinsic in a stock of physical capital varies according to whether it is foreign or locally owned. How trade patterns may be expected to change, if at all, when account is taken of variations in ownership, I have not investigated.

Policy discussions heretofore have debated whether even to condone South-South trade, let alone to cradle it, under the assumption of its relative capital intensity. A policy of positive support, however, deserves serious consideration now that its skill intensity has been demonstrated.

A recognition of the skill intensity of South-South trade also compels a new construction of the term "intermediate technology," whose traditional connotation has been a middle ground of whatever sort: technology that is neither overly labor nor capital intensive; that is neither ultramodern nor backward; that is produced by medium-size firms, etc. In a three-factor world, of course, midway techniques in the old usage are meaningless but, more significantly, the technology underlying South-South trade may be intermediate only insofar as it tends neither to be the easiest nor the most difficult to transfer. Other dimensions of skill-intensive production may conceivably involve the latest technology and the largest firms.

Finally, whether the *NICs* succeed in carving a new niche for themselves in the inter-

national division of labor will depend on technological changes in the skilled trades in the industrialized countries. The impending microelectronics revolution may squelch the *NICs'* ambitions by raising the productivity of small batch production and by reducing the demand for skilled production labor. Agnosticism about this likelihood, however, is warranted: ever since the time of F. W. Taylor, we have been led to believe that the de-skilling of such industries is just around the corner.

### REFERENCES

- Amsden, A. H., "The Industry Characteristics of Intra-Third World Trade in Manufactures," *Economic Development and Cultural Change*, October 1980, 29, 1-19.
- Balassa, Bela, "A 'Stages Approach' to Comparative Advantage," Staff Working Paper No. 256, World Bank, 1977.
- Braverman, H. *Labor and Monopoly Capital*, New York: Monthly Review Press, 1974.
- Lary, H. B., *Imports of Manufactures From Less Developed Countries*, New York 1968.
- Lewis, W. Arthur, "A Review of Economic Development," *American Economic Review Proceedings*, May 1965, 55, 1-16.
- Nelson, Richard R. and Winter, Sidney G., "In Search of a Useful Theory of Innovation," in K. A. Stroetmann, ed., *Innovation Economic Change and Technological Policies*, 1977.
- Saunders, C. T., *Engineering in Britain, W. Germany, and France*, Brighton: Sussex European Economic Research Center, 1978.
- Stern, R. M., "Capital-Skill Complementarity and US Trade in Manufactures," in H. Glejser, ed. *Quantitative Studies of International Economic Relations*, North-Holland 1976.
- Teitel, Simon, "Productivity, Mechanization and Skills," *World Development*, 1981, 9, 355-71.
- United Nations Industrial Development Organization (UNIDO), *World Industry Since 1960*, New York 1979.

# Linking Up to Distant Markets: South to North Exports of Manufactured Consumer Goods

By DONALD B. KEESING\*

Finished consumer goods have emerged as the most important category of manufactured goods exported from developing to developed countries; they now make up at least half the total. This is a trade in products ready to be sold in stores on the other side of the world. The products exported are mostly labor intensive, ranging from clothes and shoes to transistor radios, digital watches, or the latest television games.

Trade in finished consumer goods has grown astonishingly. From 1970 to 1980, OECD imports of goods in categories consisting mainly of manufactured consumer goods increased in nominal value 14.55 times while total manufactured imports from developing countries increased 10.84 times. (These comparisons exclude Australia and New Zealand which were not OECD members in 1970.) Using the UN index of the unit value of developed country exports of manufactured goods as a deflator, the "real" growth rate for all manufactured goods was 14.0 percent a year over the decade. For finished consumer goods, real growth averaged 17.4 percent. For finished consumer goods other than apparel, the growth rate averaged 20.4 percent a year. In apparel, where export opportunities were increasingly curtailed by quotas, growth averaged 15 percent a year.

The most successful developing economies in exporting finished consumer goods have been Hong Kong, Taiwan (China), and the Republic of Korea. As Table 1 shows, in 1980 these three economies exported slightly over half of the manufactured goods shipped from developing economies to OECD coun-

tries. They enjoyed a larger share of the market for sixteen categories of consumer goods—close to 72 percent. Other Asian countries supplied another 19 percent, while 7 percent came from Latin America and the Caribbean.

The United States is the leading market for these same sixteen categories, buying 46 percent, compared to 14 percent shipped to Germany, 8.5 percent to the United Kingdom, and 36 percent to the European Community as a whole in 1980.

Most of these finished consumer goods are made by developing economy firms—enterprises started, owned, and managed by people from the developing economies, usually entrepreneurs who started small, often quite recently. Such firms are especially dominant in apparel, footwear, and other simple products. As Gerald Helleiner's research on "related party transactions" in the United States helps to show, large shares of products such as radios, watches, television sets, cameras, or passenger cars are made by developed country "multinationals." On an overall basis, however, at least 80 percent of all finished consumer goods appear to be made by developing economy firms.

Developing economy firms have thus succeeded remarkably in linking up their production with the needs and fast-changing demands of stores and customers thousands of miles away. In research at the World Bank, through interviews and consultant papers, we have been exploring the institutional arrangements and the marketing aspects of this trade to learn how this linking is achieved. Here in sketch form are a few of the findings.

Although some developing economy manufacturers distribute their own products (more on this later), most exports of finished consumer goods are made to buyers' orders. Production is not begun until the exporting

\*The World Bank. The views expressed in this paper are my own, and not those of the World Bank which, however, supported the research under project RPO 671-68, "Key Institutions and Expansion of Manufactured Exports." Principal consultants were Lawrence H. Wortzel and Camilo Jaramillo.

TABLE 1—EXPORTS TO OECD IN SIXTEEN CATEGORIES CONSISTING MAINLY OF CONSUMER GOODS, BY REGION OF ORIGIN, 1980

	Hong Kong, Taiwan (China), and Republic of Korea	Other East, Southeast or South Asia	Latin America and Caribbean	Africa, Middle East, and Oceania
Sixteen Categories as Percent of Region's Manufactured Exports to OECD	61.6	32.3	18.0	15.6
Region's Percent Share of Developing Economy Exports to OECD				
Fifteen categories other than apparel	77.4	14.1	7.7	0.8
Apparel	65.9	23.9	5.8	4.4
Sixteen categories of mainly consumer goods	71.7	19.0	6.8	2.6
Other manufactured goods	34.6	30.8	23.9	10.7
Total manufactured goods	50.7	25.7	16.4	7.2

Source: Computed from OECD, *Trade by Commodities, Market Summaries: Imports, January–December, 1980*.

enterprise receives an order complete with a letter of credit or equivalent commitment to pay for the goods. The buyer specifies in full detail the design of the product, the materials to be used, the numbers and sizes to be made, and such other matters as the way the product will be labeled, packed and shipped. Designs and requirements change from order to order. In many of the same industries, production to buyers' orders is common also in developed countries. Developing economy firms are expected to achieve comparable standards.

Manufactured consumer goods are generally shipped already packaged and labeled, ready to be put straight onto the retailer's counter or clothesrack, or handed in a box to the customer. (Garments are often shipped in a specified assortment of sizes and colors; furniture is usually shipped knocked-down to be assembled in the retail store.) The developing economy manufacturing enterprise is expected to put together the entire package complete with documentation. Buyers care about all the details, including, for example, the accessories, printed labels, and packaging.

The buyer is responsible for marketing and distribution in the country of destination. Thus he takes the risks that the product will not sell as hoped. The buyer and the exporting enterprise each take risks that the product will not be made on schedule, or will

have to be rejected because of defects when inspected prior to packing, or will prove defective when sold to retail customers. The buyer risks losing customers and business while the exporting enterprise risks not being paid and not getting further orders.

The most important categories of buyers are importers and retailers. Firms known almost universally as importers are really importer-wholesalers. They tend to specialize in a rather narrow range of products. The bigger importers buy in large volume, often in several continents. To supply retailers they carry a large inventory based on their own (not necessarily original) designs. Their success depends on anticipating market trends and, at the same time, holding down costs wherever possible, not least by seeking out low-cost sources of supply.

The search for low-cost suppliers often leads importers into developing economies where the industry lacks export experience. Importers and other buyers who open up this trade commonly provide advice and teaching to exporting enterprises as necessary on practically all aspects of the business, while at first supervising every step. As the enterprise learns, a stable relationship evolves through which the importer recovers the investment in teaching. Importers also buy from experienced firms. An importer can give large orders and live with delayed delivery because of a large inventory. Inter-

views suggest, however, that importers in many lines of business bargain hard and squeeze their manufacturers' prices down as much as possible. They also tend to be quick to move on to new sources of supply or new varieties of merchandise, and will desert an established buying relationship over a small rise in costs if there is a good alternative source.

Retailers, especially large chain stores, do much of their buying directly. The proportion bought varies by product: it is high in most garments, for example, but low in footwear. As buyers, retailers generally shun direct buying from unreliable or inexperienced suppliers. They move in to cut out the middleman, however, once the exporting enterprise has become good at its tasks and requires only moderate supervision.

Retailers from different types of stores have different objectives. Some stores sell merchandise with little styling but low prices; they search for bargains or suppliers with very low costs. Others are concerned about quality, styling, and punctual delivery to fit into their merchandise plans for a coming season. These retailers are usually willing to offer prices at least a few percent higher than those paid by importers, though goods made in developing economies are often sold as low price or bargain items. Retailers known for quality merchandise seldom quibble over price, but seek to develop long-term relationships with exporting enterprises that can meet their needs.

A third important category of buyers are manufacturers or erstwhile manufacturers from developed countries, who still design their own products and market them under their established brand names, but subcontract the actual production (or part of it) to be competitive with imports. Like the better quality retailers, manufacturer-buyers are generally concerned to achieve reliable quality and stable relationships.

Practically all buyers including retailers and manufacturers are prepared to help experienced enterprises make new and difficult orders; thus supervision is given for adjustments in production to make a new design. In addition, buyers provide much useful advice and information in placing orders and

in their routine visits to check quality control and to inspect the product.

Eventually exporting enterprises become adept at meeting buyers' needs. The learning process is usually far from easy, however, even with advice and supervision from buyers. During this process, inexperienced enterprises find themselves exposed to large risks, not least financial, and survive precariously; many suffer losses from mistakes or failures; many must invest added funds to keep going; some must look for new buyers.

The enterprises that survive acquire experience in manufacturing a variety of designs, and in some cases a succession of products. Learning to make entirely new items is sometimes necessary for survival in face of changing demand. Some enterprises make profits, expand, and diversify into additional products as they gain experience.

Skills are gained that are attractive to buyers. Many an enterprise comes to know exactly what is required in exporting to each of its major markets, for example, styling and sizes and documentation, and is ready to put together the whole package reliably with a minimum of instructions and assistance. Some entrepreneurs acquire the ability to study a sample of a product, or a set of specifications, and quickly translate it into a suitable production process using equipment and labor on hand in the enterprise, perhaps not only to make the product but to "knock off" copies quickly. Some enterprises learn to make high-priced, high-quality versions of products, using more expensive materials and designs. Some become expert in quality control.

Exporting enterprises also acquire valuable experience in the marketing part of their business. This involves attracting or finding suitable buyers and inducing them to place orders. Fundamentally, the enterprise tries to sell its production capacity and capabilities. By satisfying buyers, it builds a regular business with them, which may lead it to have full order books and a roster of regular customers.

Attracting buyers is made simpler by the circumstance that each category of buyers puts much effort into finding promising suppliers in economies known for low costs and

export-oriented policies. Buyers try to learn what they can about the capabilities of any enterprise that might serve as a supplier, including what it has been making and for whom. Even so, a new buyer typically comes to visit the enterprise before placing an order. Usually the management welcomes visitors in a showroom displaying attractive items the firm has made. A new buyer is likely to want to see the plant in operation as well. In his visit a buyer seeks to assess the firm's management; when he discusses what he wants made and how, he must be reassured that the firm can do the job. The firm in turn must decide which orders to accept; usually it has a minimum size limit.

Exporting enterprises frequently take the initiative in contacting buyers and trying to interest them in placing orders. Often this requires travel. Many of the managers interviewed traveled at least once a year to major developed country markets, visiting actual and potential buyers, while enriching their knowledge of the business in other ways as well.

Managers and sales personnel ordinarily travel overseas with samples of items their firm has made. Frequently these are based on designs made for other buyers—usually a design is exclusive only for the one market so that it can be offered to customers elsewhere. In other cases, designs are copied from samples bought abroad or from pictures in magazines. Buyers nevertheless appear to judge firms partly by their sense of what is an attractive design or the latest style, made well. Some enterprises producing mainly to order send their designers abroad to improve their skills. The most experienced and successful firms in various products offer a choice of styling to their customers, and play an active role in helping the buyer select a design. However, developing economy firms that specialize in design—as do many small Hong Kong garment firms—seem to enjoy little financial success.

Together with firms that manufacture for export, exporting enterprises also include trading companies that organize production for export by smaller firms or handicraft workers. A buyer deals mainly with the trading company, though he may be allowed to

supervise its subcontractor. Large trading companies, which are now being promoted in many developing economies by special incentives, help to provide active marketing abroad, maintain offices there, and approach wholesalers and retailers in search of orders.

Some exports from developing countries are made, not to buyers' orders, but on the basis of production for inventory with sale by manufacturers' representatives—usually nationals of the importing country, working on commission or directly for the developing economy manufacturer. This sort of distribution is limited, as a rule, to large firms making standardized products in only a few basic designs and sizes.

Products distributed directly by firms from Brazil included towels, T-shirts, jeans, and men's robes in cotton terrycloth. Furniture manufacturers from more than one country have been found to have their own distribution subsidiaries. Some of the television set manufacturers from Taiwan (China) and Korea have begun to sell through American sales representatives, although with only limited success. Distribution by manufacturers has also been encountered exceptionally in other products, including radios, watches, and electric fans. Most of the products distributed in this way are also commonly made to buyers' orders.

One large Korean firm has its own wholesale distribution in the United States, even though its products—mainly shirts and other garments, but also leather shoes—come in many designs and sizes. When interviewed in 1978, from its warehouses in New Jersey and Los Angeles it invoiced to 600 stores. In shirts it offered customers a choice of 40 patterns, though much of its output (6 million shirts per month) continued to be made to buyers' orders. The firm promoted by having 1,500–2,500 people travel each year to visit customers.

No developing economy brand or product line was being advertised to households or individual consumers in developed countries, at the time of our East Asian interviews. All sales by manufacturers through their representatives or distributors were made to retailers or wholesalers in the country of destination, frequently with a choice of brand

names. However, many developing economy firms vigorously promoted their own brands and product lines at home and, in some cases, in other developing economies.

For developing countries trying to export, our findings confirm the importance of measures that improve incentives and reduce costs. Buyers know what prices they can pay and are ready to switch to competing sources of supply elsewhere in the world. Thus only where exchange rates and policies bring about low costs—above all of labor and high quality materials—will buyers place orders and teach new suppliers. Orders are given by preference in economies with easy, duty-free access to imported inputs, since this makes it much easier to get together on time a complete, packaged product and to move on to new designs or materials. Experience has shown that much tends to go wrong when exports are ordered in economies with inward-looking policies: export orders are neglected or refused because of higher returns

in the domestic market, inputs are not delivered, costs rise and cease to be competitive, etc.

These findings are also a reminder that export responses to policies depend on people, and thus on the information conveyed and expectations created. Policy changes begin to be successful when they cause buyers to change their perceptions and shift their search to a new country. Results also depend on persuading entrepreneurs, local or foreign, to try to export from the country. Since buyers and entrepreneurs do not think exactly like economists, we may not be the best source of advice on their responses to policy measures.

#### REFERENCE

- Helleiner, Gerald K., *Intra-firm Trade and the Developing Countries*, New York: St. Martin's Press, 1981.

# New Theories of Trade Among Industrial Countries

By PAUL KRUGMAN\*

Most students of international trade have long had at least a sneaking suspicion that conventional models of comparative advantage do not give an adequate account of world trade. This is especially true of trade in manufactured goods. Both at the macro level of aggregate trade flows and at the micro level of market structure and technology, it is hard to reconcile what we see in manufactures trade with the assumptions of standard trade theory.

In particular, much of the world's trade in manufactures is trade between industrial countries with similar relative factor endowments; furthermore, much of the trade between these countries involves two-way exchanges of goods produced with similar factor proportions. Where is the source of comparative advantage?

Furthermore, most manufacturing industries are characterized by at least some degree of increasing returns (especially if we include dynamic scale economies associated with *R&D* and the learning curve). Not coincidentally, most manufacturing industries are also imperfectly competitive to at least some extent. Can a model which assumes constant returns, exogenous technology, and perfect competition give adequate guidance for trade policy in these industries?

In response to these questions, many economists have proposed alternatives to conventional trade theory. The alternatives include the "product cycle" view, with the stress on endogenous innovation and the diffusion of technology; the arguments of many observers that much trade among industrial countries is based on scale economies rather than comparative advantage; and the common argument that a protected home market can promote exports. Until recently, however, none of these alternatives was presented in a form which economists would properly call a model: that is, a formal struc-

ture in which macro behavior is derived from micro motives. This lack of formalization essentially barred alternatives to comparative advantage, however plausible, from the mainstream of international economics.

In the last five years or so, however, there has been a significant change. A number of theorists have begun to apply methods drawn from the theory of industrial organization to international trade, to produce a new genre of trade models. These models offer a new way of looking at trade—and particularly at manufactures trade among the industrial countries.

A characteristic feature of the new models is that they often rely on very special assumptions. This is probably inevitable: given the inherent complexity of the world once the great simplifying device of constant returns is dropped, only special assumptions will yield tractable analysis. In spite of the specialness of individual models, however, the new literature on trade is starting to give rise to concepts which look more general than the particular models used to illustrate them. The purpose of this paper is to sketch out two such concepts which I believe are important and more general in application than the particular models in which they have been expressed. The first is the theory of "intraindustry" trade, a view which incorporates scale economies as well as comparative advantage as major causes of trade and gains from trade. The second is the (less well developed) theory of technological competition, which may begin to shed some light on the dynamics of international competition in research-intensive industries.

## I. The Theory of Intraindustry Trade

It has long been known as a theoretical point that increasing returns can be an alternative to comparative advantage for the explanation of trade. It has also been suspected by many economists that scale economies do

\*Massachusetts Institute of Technology.

in fact play a major role in manufactures trade among the industrial countries—perhaps more important than differences in factor endowments. The problem in making this more than just a wise remark has been the difficulty of introducing scale economies into formal models of trade.

The traditional way of doing this is to assume that increasing returns are wholly external to firms. The models which result from this assumption, however, have never had much influence. External economies are too vague and unmeasurable to be an appealing explanation of trade patterns. To have the right “feel,” it appears, a formalization of the role of scale economies must lay its stress on internal economies of scale—and this, until recently, was not something trade theorists knew how to do.

In the last few years, however, a relatively coherent view of the role of scale economies in trade has finally emerged. This view—which we might rather grandly call the “theory of intraindustry trade”—was developed by a number of authors who found in recent developments in monopolistic competition theory the modelling techniques needed.

The basic idea of the theory is extremely simple. We distinguish between two kinds of trade: interindustry trade based on comparative advantage, and intraindustry trade based on economies of scale. The *industrial* structure of a country's production will be determined by its factor endowments. Within each industry, however, there is assumed to be a wide range of potential products, each produced under conditions of increasing returns. Because of these scale economies, each country will produce only a limited subset of the products in each industry, with the pattern of *intraindustrial* specialization—which country produces what—essentially arbitrary.

The implications for the trade pattern are straightforward and empirically plausible. Each country will be a *net* exporter in industries in which it has a comparative advantage, just as conventional theory suggests. Because of intraindustry specialization, however, each country will import some products even in industries in which it is a net exporter, and vice versa; that is, there will be

intraindustry as well as interindustry trade. Furthermore, the more similar countries are in their factor endowments, the less different their industrial structures will be, and hence the more their trade will have an intraindustry character.

If the theory of intraindustry trade is so simple, why is it a new development? The answer is in part that it is not: the basic story just described may be found in many informal discussions of trade in the 1960's. A formal model of intraindustry trade, however, must deal with the problem of market structure. The existence of unexhausted economies of scale means that markets cannot be perfectly competitive. They could, however, be characterized by Chamberlinian monopolistic competition, and in fact the product differentiation-cum-scale economies story seems to dovetail very naturally with this approach. Notice that we need not believe that Chamberlinian equilibrium is actually a realistic description of the world. The point is that it is a useful *device* for closing the model, and is in some sense less unrealistic in this context than perfect competition. Thus a number of economists, including Avinash Dixit and Victor Norman, Kelvin Lancaster, Elhanan Helpman, Wilfred Ethier, and myself (1981) have presented Chamberlinian models which formalize the story just described. These models differ in detail, but bear a strong family resemblance to one another, justifying us in referring to a “theory.”

The theory of intraindustry trade, then, provides a neat explanation of the empirical puzzles posed by manufactures trade among the industrial countries. It explains both why similar countries trade so much, and why so much of their trade is two-way exchanges of similar products. It also provides some interesting new insights into the effects of trade on welfare and income distribution. Traditional models have very strong distributional effects: even though trade liberalization may potentially make everyone better off, the movement of the income distribution is always enough to insure that the real income of scarce factors of production falls. If there are increasing returns to scale, however, this need not be the case. Scale economies pro-

duce some extra gains, in the form of longer production runs and a greater variety of products. If changes in the income distribution are not too large—if, for example, trade liberalization takes place between countries with similar relative factor endowments—the advantages of a larger market can outweigh the distributional effects of trade. This means that the distributional effects of trade may depend on its causes. If scale economies are relatively unimportant and countries differ substantially in factor endowments, we have the conventional Stolper-Samuelson result that scarce factors lose from trade. If, on the other hand, scale economies are important and factor endowments are similar, all factors gain from trade.

But when would these conditions hold? The situation of significant scale economies combined with weak comparative advantage is precisely that of trade in manufactured goods among industrial countries. If we believe that trade liberalization is easiest when nobody gets hurt, this may help explain why the great trade liberalization of the postwar period has focused on manufactures trade between advanced nations.

## II. The Theory of Technological Competition

The theory of intraindustry trade seems to suggest that trade in manufactured goods among industrial countries is a benign thing, less likely to cause adjustment problems than other trade, and hence easier to liberalize. Another strand in recent theory, however, suggests that in some manufacturing sectors—specifically, those where *R&D* play a crucial role—there may be a strong temptation for countries to engage in protectionist or interventionist policies.

Here again the basic concept is quite simple. I adopt a more partial view than in the last section, focusing on a single industry; I assume that in the industry there are only two firms, one domestic and one foreign. Suppose these firms can compete technologically, investing in *R&D* to lower their costs, develop new products, or both. The amount they spend on *R&D* determines their position in a later competition in actual product markets. It is possible to envision a variety of

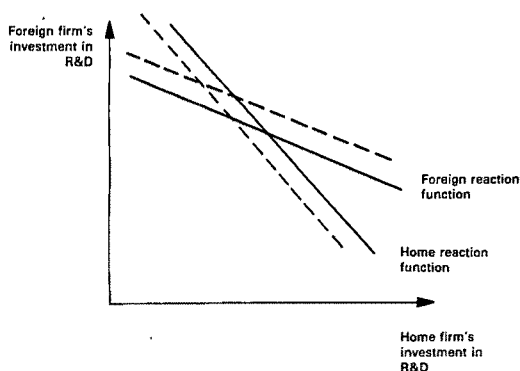


FIGURE 1

particular models of this type: the *R&D* may apply to product or process improvement, it may be certain or uncertain in its results, product development may have a winner-take-all aspect or leave room for second prizes, and so on. But for a wide variety of particular models, the basic technological competition will be summarizeable by a diagram like Figure 1. Each firm's optimal investment in *R&D* will be declining in the other's investment, as indicated by the two reaction functions; the noncooperative equilibrium will be where the schedules cross.

I should note as an aside that a similar analysis might also be applied where technology is improved, not through formal *R&D*, but through learning by doing. The technological competition would then take the form of willingness to accept low initial earnings to move faster down the learning curve; the qualitative character of the results would probably be much the same.

As in the case of intraindustry trade models, the simplicity here is in some respects misleading. To derive these schedules explicitly requires analysis of a kind that has only recently become well understood, through the work of such authors as Dixit. We must solve the model backwards: first deriving the post-*R&D* equilibrium conditional on the levels of *R&D* expenditure, then using the results of this analysis to derive the relationship of expected profits to *R&D*, which allows us to draw the reaction functions. Thus behind the simplicity of the figure lies a complex competitive process. I

have drawn the diagram with "nice" properties: the curves slope downwards, there is a unique equilibrium, and it will be stable under plausible adjustment schemes. To show that the curves actually have these properties, or to find the restrictions on parameters necessary to insure that they do, will be far from trivial.

Suppose, however, that we assume that technological competition in some industries can actually be reasonably well represented in this way. We can then use the diagram to think about a subject which is central to debate over trade policy but virtually untouched by trade theorists. This is the effect of trade policy on technology. What I have in mind, in particular, is the argument that a protected home market can give a country's high technology firms an advantage which eventually gives them an edge in export markets as well.

It seems clear from Figure 1 that this argument does make a good deal of sense. Suppose that the foreign government denies the home firm access to part or all of its market. What this denial of access will do in a variety of particular models is to raise the expected return to a marginal dollar of *R&D* by the foreign firm, lower the expected marginal return to *R&D* by the domestic firm. Thus the reaction functions will shift in the directions indicated by the dotted lines. Foreign *R&D* will be greater than it would otherwise have been; domestic *R&D* will be less. Because the foreign firm's relative technological position is improved, it may well increase its share of unprotected as well as protected markets. In other words, import protection will turn out to be a form of export promotion.

Is this desirable from the point of view of the foreign government? Recent work by Barbara Spencer and James Brander suggests that it may be, although *R&D* subsidy is probably a better policy. As they point out, in imperfectly competitive markets there is some monopoly rent for which firms are competing. Government action may enable domestic firms to seize a larger share of these rents than they would otherwise be able to get.

Introducing technological competition into trade theory, then, does seem to give some justification for the kinds of industrial policies which Japan is accused of following. Or at any rate, it offers support for the idea that protecting *R&D*-intensive industries may really be a beggar-thy-neighbor policy, not simply a beggar-thyself policy, which conventional theory would suggest. This by no means clinches the case for protectionism, but it gives some reason to be more worried about foreign targeting of high technology industries than about other trade-distorting practices.

This paper has sketched out two new approaches to trade among industrial countries, based on the recent emergence of a literature which applies concepts from industrial organization theory to international trade. Of the two, the theory of intraindustry trade is a relatively finished product, while the theory of technological competition is still in a rough state. Both are, I hope, of some use for thinking about issues—including important policy issues—which cannot be handled by traditional theory.

## REFERENCES

- Brander, James and Spencer, Barbara, "Strategic Commitment with *R&D*: The Symmetric Case," mimeo., 1982.
- Dixit, Avinash, "The Role of Investment in Entry-Deterrence," *Economic Journal*, March 1980, 90, 95-106.
- and Norman, Victor, *Theory of International Trade*, Oxford University Press, 1980.
- Ethier, Wilfred, "National and International Return to Scale in the Modern Theory of International Trade," *American Economic Review*, June 1982, 72, 389-405.
- Helpman, Elhanan, "International Trade Under Economies of Scale and Imperfect Competition: A Chamberlin-Heckscher-Ohlin Model," *Journal of International Economics*, August 1981, 11, 305-40.
- Krugman, Paul, "Intraindustry Specialization and the Gains from Trade," *Journal of Political Economy*, October 1981, 89, 959-73.

\_\_\_\_\_, "Import Protection as Export Promotion," in Henrik Kierzkowski, ed., *Monopolistic Competition in International Trade*, Oxford University Press, forthcoming.

Lancaster, Kelvin, "Intra-Industry Trade Un-

der Perfect Monopolistic Competition, *Journal of International Economics*, May 1980, 10, 151-75.

Spencer, Barbara and Brander, James, "International R&D Rivalry and Industrial Strategy," mimeo., 1982.



## THE IMF AND CONDITIONALITY

### Lender of Early Resort: The IMF and the Poorest

By G. K. HELLEINER\*

Liquidity is required both for countries experiencing temporary balance of payments difficulties and for those seeking smooth adjustment to a permanently changed external or other environment. In current circumstances, most of the IMF's poorest members are facing both types of liquidity requirement at the same time. The severe world recession with its disastrous impact upon their terms of trade must be regarded as a temporary phenomenon; but, even prior to its full onset, these countries faced the need for adjustment to permanent oil-price-related deteriorations in their external terms of trade. Both temporary and permanent shocks have generated further domestic macroeconomic maladjustments, reflected most notably in inflation and rising real effective currency values.

Before plunging into the hot current debate about the design of stabilization or adjustment programs, one ought to address the question of "uniformity of treatment" with respect to short-term balance of payments financing for the various members of the IMF. This is quite apart from, and additional to, the still unresolved issue of symmetry in the bearing of adjustment burdens as between surplus and reserve currency countries and the rest. Uniformity of treatment requires that account be taken of: (i) the differential availability of fast-disbursing short-term balance of payments financing (liquidity); (ii) differences in the size of shocks to the external account; and (iii) differential capacity to adjust.

#### I. Availability of Liquidity

Liquidity derives today not only from reserves of gold and foreign exchange, but also

from borrowing possibilities which can be relied upon with a minimum of delay and conditionality. The poorest countries benefited scarcely at all from the two principal new sources of international liquidity in the 1970's: the substantial increase in the price of gold (since they held very little); and the expansion of commercial bank overseas lending. Nor are they likely to benefit from emergency BIS finance such as was found for Mexico.

Commercial bank credits are offered to the poorest only to a very limited extent. A few poor countries have resources and longer-term potential which have led banks to consider them acceptable credit risks, for example, China, India, Zaire. In mid-1981, according to BIS data, low-income countries had short-term bank credits outstanding (one year and under) totalling about 8 percent of the total value of their trade. Medium- and longer-term commercial finance is also frequently to be found in these countries, but it too is limited in amount and frequently associated with particular export contracts rather than freely available to the borrower. These commercial credits are all offered at rates of interest that include risk premia which are well above the developing country average.

An increasingly important source of low-conditionality balance of payments credit for the majority of the poorest countries has been arrears on current commercial payments. At year-end 1981, there were 32 countries reported by the IMF as in arrears, of which half had per capita incomes of under \$410 (in 1980). Apart from the fact that these arrears are likely to carry very high rates of interest and to interrupt the normal flow of trade, this source of (forced) external finance only gains the debtor a once-and-for-all and very temporary advantage; for it leads to subsequent reduction in credit which would otherwise have been enjoyed.

\*Department of economics, University of Toronto.

With neither gold holdings nor significant access to commercial credit, the poorest have had only two means of expanding their liquidity. The first was to build up their own earned foreign exchange reserves. In the aggregate, the low-income countries did rebuild their reserves in the second half of the 1970s, although, excluding China and India, they never regained the reserve/import ratios of the early 1970's. The opportunity cost of holding reserves in the poorest countries, those most requiring increased real inputs to development efforts, is high. Few would have recommended levels of reserves great enough to have permitted the stabilisation of imports through a recession of the current one's severity.

The second source of increased liquidity for the poorest has been increases in drawing rights in the IMF. The IMF quotas roughly doubled between the beginning of 1971 and the present. This implies a doubling in access to low-conditionality IMF credit—in the first credit tranche of the general account, and the first 50 percent of quota under the compensatory financing facility (*CFF*). The second issue of special drawing rights (*SDR*) in 1979–81 together with the cessation of their reconstitution provisions added further such finance, making a total increase over the decade of about 120 percent in maximum annual access to low-conditionality IMF credit.

The trade of the poorest countries grew at slower rates than that of other developing countries, but it still expanded at substantially more rapid rates than did these countries' access to low-conditionality IMF credit. Between 1971 and 1981, according to UNCTAD data, imports grew by 341 percent in the least developed countries and by 426 percent in countries of under \$500 per capita; exports grew by 183 percent in the least developed and by 399 percent in the under \$500 per capita income category. Thus, in a period of substantial general increases in the instability of the terms of trade, when other countries were benefitting from other new sources of expanded liquidity, the poorest countries' principal source of unearned liquidity expanded at a rate significantly less than that of the value of their trade.

Conditional borrowing rights in the IMF expanded over the 1971–81 period, in consequence of the policy of "enlarged access" and expansion of the *CFF*, at considerably more rapid rates. Total IMF annual credit availability, as distinct from low-conditionality credit, has expanded over the past decade by about 350 percent. This rate of expansion in annual credit availability is of the same order as the rate at which the trade of the poorest grew in the 1970's, though arguably, less than the increased instability of these countries now required.

Whereas in the 1974–75 period, two-thirds of IMF credit was granted without significant conditions, much of it via the temporary oil facility, in the last two years only 20 percent was granted in this way. This increase in conditionality together with a recent increase in the stringency of IMF conditions—measured by such indicators as length of programs, number of monthly credits, and degree of insistence upon exchange rate action—should be proof against the charge from some quarters that the IMF has been "too soft" in its lending to developing countries. So stringent have IMF conditions become that in the first half of 1982, a period in which the terms of trade of primary exporting countries were the worst in over thirty years, the IMF cancelled more of its previous commitments than it made new ones. Even gross new commitments (of all kinds) were running at about one-fifth the rate of the first six months of the previous year.

The difference in the "liquidity" of high- and low-conditionality finance is not always adequately appreciated. Liquidity implies fast disbursement and flexibility of use. Conditional IMF credits, on the contrary, can involve very high transactions costs. Detailed negotiations over preconditions and performance targets and the continuing uncertainty associated with the quarterly review process extract a heavy cost from the borrowing country in terms of the time and effort of its most able policymakers. The opportunity cost of this time and effort is particularly high in countries where, as in the poorest, skilled manpower is especially scarce, and agreement with the IMF does not bring further external finance from commercial sources.

The slower growth of low-conditionality finance than imports in the poorest countries during the 1970's would alone imply, other things being equal, a decline in their liquidity. In fact, their situation is considerably more serious, as indicated by their depleted reserves, arrears, and increasing difficulty in acquiring conditional credit from the IMF.

## II. Size of External Shocks

There are many ways in which the dimensions of external shocks or the degree of external instability can be described. For the present purposes, instability has been measured as the coefficient of variation (standard error divided by the mean) of a linear time trend. (Data available on request.) Developing countries of all (UN) categories have experienced greater instability in this sense, both in commodity terms of trade and in purchasing power of their exports (income terms of trade), than the developed market economies in the 1960's and 1970's. The least developed countries (UN definition) have been and still are, on average, the most unstable of the developing country groups with respect to their commodity terms of trade, although "fast-growing exporters of manufactures" and the developed market economies experienced greater *increases* in instability from the 1960's to the 1970's. Very large increases in the instability of the least developed countries' income terms of trade in the 1970's made them the most unstable in this dimension as well.

Import volume, which is the key measure of the *total* degree to which domestic stability is hit by external events (or, in some instances, the degree to which domestic events are themselves unstable), is also more unstable in the developing countries than in the developed market economies. The least developed countries' increased real export instability is fully reflected in their import volume instability which has become higher, on average, than that of all other non-oil developing countries in the 1970's. Most interesting is the *decline* in measured import volume instability in the 1970's for the "top" categories of developing countries—fast growing exporters of manufactures, and those with per capita income of over \$1,000—de-

spite substantial increases in the instability of their commodity and income terms of trade. This undoubtedly reflects the improved access which these countries acquired in the 1970's to short- and medium-term commercial finance.

Some of the poorest countries were able somewhat to smooth their import volume with increased external private credit in the same way as the slightly richer ones did. Others benefitted from increasing, and more stable, earnings on services account. In some, an element of stability was also imparted to the volume of imports by the fact that high proportions were paid for out of official development assistance rather than export earnings.

In the 1980–82 period, the commodity terms of trade of the low-income countries dropped on average by some 25 percent to levels lower than any in thirty years. The purchasing power of these countries' exports plunged nearly as much. Adjustment to these external events was most severe in the low-income countries where, on average, there were declines in the volume of imports, averaging roughly 2 percent in 1980, 7 percent in 1981, and probably at least as much in 1982. These import cuts were necessitated by these countries' inability to finance current account deficits even of previous dimensions. In 1981, the low-income countries actually received less finance than in 1979, whereas there had been substantial expansion in the finance made available to other developing countries. The already perilously low levels of reserves of the weakest of these countries were at the same time further reduced.

## III. Capacity to Adjust

The missing element in the IMF's application of "uniform" conditionality to its members is consideration of their differential *capacity to adjust*, more particularly, the limited capacity of the poorest or least developed to benefit from stabilization programs such as the IMF typically recommends.

"Capacity to adjust" involves many dimensions: (i) the structure of the economy—particularly the relative size, degree of concentration, and composition of the export

sector; (ii) the degree of economic rigidity, particularly in terms of price responsiveness and fiscal flexibility of the economy; (iii) the per capita income and/or the real level of urban wages, and recent trends therein; (iv) the technical and administrative capacity of the economic policy making arm of the government; and (v) the degree of social and political consensus behind the government.

With respect to all but the last, it is safe to say that the least developed and the majority of the poorest countries are always at a significant disadvantage, relative to other countries, in their capacity to achieve rapid adjustment to external shocks which hit all countries with equivalent terms of trade effects.

The typical poor country has a relatively large, primary product-based and concentrated export sector. Exports of minerals or such agricultural products as tropical beverages or hard fibres (the staples of the export sector) are not substitutes in domestic consumption. Nor is it typically possible quickly to deflect articles of local consumption into exporting, if only because the required marketing infrastructure is not there. Its imports are usually made up of consumer essentials, intermediate inputs, and capital goods, demands for all of which are fairly price inelastic. "Getting prices right" is slow-acting medicine in the poorest countries where markets function more imperfectly because of rigidities, inflexibilities, and market segmentation. Governments are usually highly dependent for their revenues upon the external sector, and have few alternative sources. From these structural characteristics and this degree of economic rigidity, it follows that terms of trade shocks will register enormous impacts upon the economy and government revenues, larger than in more developed economies; and that when such shocks are permanent, they cannot quickly be accommodated via supply-side readjustments. The short- to medium-term burden of adjustment inevitably falls disproportionately upon income and the volume of imports.

Per capita income and urban real wages, when at levels already the lowest anywhere, are likely to be difficult to squeeze further in pursuit of "adjustment" programs. In many

of the poorest countries, per capita income growth has already been stagnant or declining for some time. In the past two years, cuts in levels of consumption and investment have been sharper in the poorest than in other countries. Further cuts may be politically unsustainable.

It is difficult to produce statistics which demonstrate the limited technical and administrative capacities of the poorest, but there can be no doubt of these countries' problems in this respect as well. Statistical services are typically weak, the supplies of available manpower for economic policy-making are limited, the literature on macro-economic management for such countries is almost nonexistent, and when skilled manpower in these areas is imported, it is frequently not of the highest quality. It is not unusual for very poor countries to lack *any* series, however crude, on such crucial economic variables as import prices, import volumes, and nominal or real effective exchange rates. When major shocks occur in such economies, there are normally longer lags than in better-off countries between the events and domestic recognition of their implications, and even longer relative lags between such recognition and the formulation of policies to deal with them. It is easy for such governments to be overwhelmed by the effects of large shocks. By the time that the need for policy adjustment is recognized, its size may already have become more daunting than would have been the case had a stronger policymaking structure been able to respond earlier. In these circumstances, "coming early" to the Fund may not be within the range of options available to a poor country.

#### IV. Conclusions

It is anomalous and unfair that the international monetary system now expands its liquidity in such a way as to provide the least for those countries experiencing the greatest external shocks and possessing the least capacity to adjust to them. If there is inadequate liquidity, adjustment may be necessary even to temporary disturbances which should not require it. The design of adjustment programs, like the need for adjustment itself,

is heavily influenced by the availability of finance. The less the available external financing, the shorter the period during which necessary adjustment must take place; and the greater the likelihood of "shock-type" rather than more gradual adjustment programs. The poorest countries are those in which adjustment is likely to be particularly sluggish, so that the degree of shock required to achieve the desired effect may be even greater than that which the same adjustment requirements and shortages of liquidity would necessitate in more flexible economies.

Thus IMF low-conditionality credit, the only IMF finance which can be described as contributing to true liquidity, should be expanded so as to provide adequate increases in liquidity for *all* its members. This can be done in different ways, for example, through overall quota increases, *SDR* allocations, or altered provisions respecting existing borrowing rights, in each case ensuring that special provision is made for those without alternative sources of liquidity. Probably the best means for achieving the desired increase, with the minimum risk of converting the IMF into a source of long-term finance, would be the further expansion and liberalization (for example, via increased quota limits for low-conditionality finance) of the *CFF*. With current high and variable rates of global inflation (the instability of import prices rose between the 1960's and 1970's by a factor of 8 for the least developed), the *CFF* should at last be shifted to an income terms of trade base.

On the issue of the precise nature of IMF conditionality when it *is* required, one must tread cautiously. The appropriate pace and

phasing of policy change and adjustment, the distribution of the burdens, the precise mix of policy instruments, are all matters of political as well as technical judgment. Even if the IMF staff were quite exceptional in their purely technical capacities, which they often (especially in missions to smaller and poorer countries) are not, these issues are tricky and not ones on which there is a professional consensus. Getting them right requires sensitivity and appropriate experience. The IMF mission chiefs, who must try to assist in the formulation, and eventually gauge the "adequacy," of policies and programs, play a crucial role in the relationship between the IMF and its members. They can at times assist in the development of improved policies supported by international finance. But there is a major risk that conditionality can produce dubious advice, resentment, delay, and inappropriately reduced finance instead. The introduction of explicit contingency clauses into the IMF's performance criteria would ease some of the transactions costs of obtaining IMF high-conditionality credit and, in uncertain times, encourage the resumption of adequate multi-year programs instead of the current return to "short-leash" (one-year) IMF agreements. Consistent with such an innovation would be contingency clauses governing repurchase obligations as well.

## REFERENCES

- UNCTAD, *Handbook of International Trade and Development Statistics*, Washington, 1980.

# On Seeking to Improve IMF Conditionality

By JOHN WILLIAMSON\*

What is the social function served by having the International Monetary Fund (IMF) lend to its member countries? At a somewhat abstract level, I would suggest that it could be defined as that of easing the external constraint on its member countries to the extent that such an easing can be expected to be advantageous to the world community as a whole. It follows immediately that the easing has to be of the external constraint provided by *liquidity* rather than *solvency*, since at best, an easing of the solvency constraint involves a resource transfer that will leave the donors worse off, while at worst, a belief in endogeneity of the solvency constraint creates moral hazard problems that undermine the incentive for economic efficiency. Hence, my conception of the principle that should be guiding the Fund's lending policies is that they ease the external liquidity constraint to the degree that is generally advantageous while preserving the intertemporal external budget constraint dictated by solvency considerations.

## I. The Rationale for Conditionality

This principle can rationalize the parallel existence of low- and high-conditionality facilities in the Fund. There is general advantage in each country having liquidity adequate to finance temporary deficits rather than being forced to adjust to them, and in having the option of making adjustments that are needed to safeguard external solvency at a measured pace while financing the interim deficit. If liquidity adequate to fulfill those legitimate needs is provided unconditionally, there is a danger that countries with myopic governments will spend more than they should, and in the process get themselves into unsustainable deficit and the world into inflation. Low-conditionality finance could safely be provided on a more

generous scale than unconditional liquidity, provided that it were tied to the existence of objective exogenous circumstances which produce a temporary deficit. But where a deficit cannot be presumed to be temporary, then the provision of liquidity has to be conditional on adoption of a set of measures judged adequate to secure adjustment and thus ensure that the deficit will after all be temporary. That is the basic logic of high-conditionality finance.

The preceding argument suggests that low-conditionality liquidity should be provided on a rather generous scale relative to unconditional liquidity. This is not the case at the moment: on a world scale, drawing rights under the first credit tranche and the compensatory financing facility are derisory relative to reserves (even if one excludes gold and potential bank borrowing from one's concept of reserves). To change that balance by limiting unconditional liquidity would need asset settlement and international controls on bank lending, not to mention the definitive demonetization of gold—which does not sound like a promising basis for an Action Program. To change the balance by expanding low-conditionality liquidity would be redundant so far as countries with access to the international capital market are concerned. Where it could be important is in regard to the large group of low-income countries without significant access to the international capital market. For these countries an expansion and rationalization of the compensatory financing facility as urged by Sidney Dell and Roger Lawrence (1980), so as to entitle a country to borrow whenever exogenous events result in a deficit that can be presumed to be temporary, would be a very worthwhile development. It would also be logical to restructure the repayment obligations under such a facility, so that repayment fell due as the exogenous shocks went into reverse rather than on a fixed schedule. (But the low conditionality of the first credit

\*Senior fellow, Institute for International Economics.

tranche is an anachronism that cannot be justified by the rationale developed above, nor by any other of which I am aware.)

Where circumstances are such that there is no presumption that a deficit will be reversed without policy changes, the principles suggested above imply that the provision of liquidity has to be conditional on appropriate modification in policies. There may be some scope for trusting countries to introduce such policy changes of their own volition without supervision from "grandmother" Fund. For example, I have argued elsewhere (1982, p. 16) that countries faced with an exogenous but presumptively permanent payments deterioration might be allowed low-conditionality finance for a *tapering* proportion of the deficit resulting from the adverse shock. That would provide them with time to implement adjustment policies of their own choosing. But if adjustment failed to occur reasonably promptly, they would in due course be forced back into seeking high-conditionality loans—as would a country whose deficit arose from its own policy errors as soon as its unconditional liquidity was exhausted. Under such circumstances, the basic notion that the Fund should insist on adoption of a set of policy measures that can be presumed adequate to secure adjustment as a condition for granting credit is surely correct. Yet far too many of the "radical" attacks on the Fund seem to me to deny this premise, and to amount instead to demands that the Fund relax the solvency constraint on governments that the observer judges ethically meritorious.

## II. The Implementation of High Conditionality

Designation of the Fund as an "adjustment institution" whose duty may involve its withholding credit until it is satisfied that a program adequate to induce adjustment is in place requires the Fund to take a view on what constitutes an adequate adjustment program. Given the fact of national sovereignty, it would in principle be improper for the Fund to dictate the form of such a program, as opposed to satisfying itself that a country's chosen program is (with reasonable

probability) adequate to the task. There are undoubtedly occasions when the Fund has been insensitive in the way in which it has seemed to attempt to dictate a program.

But I am not persuaded that this is an issue which demands substantive changes in Fund policy, as opposed to a bit more tact by certain members of the IMF staff. If the Fund is not satisfied that the policy program initially laid before it by a country wishing to borrow will be adequate to secure adjustment, then it has to discuss how to make it adequate. In giving policy advice, it would be irresponsible of the Fund to advise only measures that would be certain to secure a payments turnaround and to ignore the costs to the country's domestic objectives (absorption, growth, the control of inflation, income distribution, etc.). It has to be concerned to devise a package that will respect the full range of objectives that countries consider important as well as achieving a payments recovery.

The more important question in my view is therefore that of identifying the type of macroeconomic strategy that will combine payments adjustment with satisfactory domestic performance. While this will vary depending on both the circumstances and the priorities of individual countries, and the Fund needs to be sensitive to variations in both of those dimensions, there is no getting away from the fact that the requirements of payments adjustment while respecting the interests of other countries imply that measures will ordinarily be drawn from a rather narrow set. Of course, any imaginative economic theorist can invent paradoxical cases in which the balance of payments would benefit by revaluation (low elasticities of demand and high elasticities of supply), credit expansion (output of exportables constrained by credit-financed imported inputs), increased government spending (on debottlenecking the tradable goods sector), or whatever; but one needs some pretty strong empirical evidence that such circumstances actually exist before one can prudently embrace the paradoxical option in a specific case. In general, Fund programs are bound to involve devaluation, credit restriction, and fiscal retrenchment.

One therefore cannot test whether the Fund exhibits proper flexibility as opposed to rigid monetarism by examining the frequency with which its programs involve devaluation or credit ceilings or cuts in public expenditure. A relevant test of whether Fund programs respond to the objective circumstances of countries is whether the recommended mix between expenditure-reducing and expenditure-switching policies is varied so as to avoid pushing a borrowing country significantly below internal balance. It is quite clear that the Fund is at times sensitive to this issue: see Hans Schmitt (1981) for an account of why the Fund urged a significant devaluation as a part of the Portuguese program of 1977. Discussion at the Institute for International Economics' conference on IMF Conditionality (the proceedings of which will appear in my 1983 volume) indicated that this was not an isolated instance. And the case studies placed before that conference yielded rather little evidence of the Fund having urged deflationary overkill, at least in the period 1976-81; in cases where that accusation was made, the interruption to growth was either slight (U.K.) or in retrospect it is clear that policy was on a completely unsustainable course (Jamaica, Turkey) or both (Peru). It is not clear, however, that examination of an earlier period would have found equally little evidence of Fund demands for overkill, nor is it clear that the tightening of conditionality in the course of 1981 has not jeopardized the enlightened attitude on that issue that had come to rule in the Fund during the period studied by the Institute's conference.

Two litmus tests of whether the Fund accepts national preferences are whether it accepts national decisions on the priority to be attached to combating inflation and whether (in the typical case where wages are above the market-clearing level in the tradable sector) the Fund offers countries the option of a higher level of real output in exchange for a cut in the real wage. On the first issue, the Fund has come a long way since 1967, when it opposed the expansionist policies that initiated the Brazilian "miracle" on the ground that reducing inflation should remain the top priority. All the evidence is

that in its relations with small borrowing countries (as opposed to the rhetoric it directs at its major members), the Fund now accepts national views on the priority to be attached to combating inflation. On the second issue, there is not much evidence that the Fund has yet faced up to the tradeoff any more realistically than its major members.

Other policies which the Fund customarily urges on member countries include maintenance of an outward orientation; it regularly requires that borrowing members undertake not to increase restrictions on trade or current-account payments. This is in part intended to protect the general international interest in a liberal order, and in part reflects an intellectual conviction that outwardly oriented policies are in the best interests of the countries that adopt them. Aside from some concern that the speed of import liberalization urged may in some cases have been precipitate (Jamaica, Peru, Turkey), the Fund's position seems to me entirely reasonable.

### III. Improvements in Conditionality

I have argued up to now that there is a logic in both the broad structure of IMF conditionality and the types of measures that feature in Fund programs. In the process I have already mentioned certain improvements that seem to me to be called for: extension and rationalization of access to low-conditionality finance where payments deficits arise as a result of exogenous shocks (at least so far as low-income countries are concerned); the restructuring of repayment obligations to tie reimbursements to objective, exogenous determinants of ability to pay; and more caution in the speed at which imports are liberalized. Several other suggestions were advanced in my study that drew on the Institute's conference (1982).

1) I argued that in addition to the traditional categories of temporary, excess demand, and fundamental deficits, it is desirable to recognize the concept of a "structural deficit." This is defined as a deficit caused by a presumably permanent adverse exogenous shock which cannot be adjusted without abandoning internal balance other than by

structural change. (The term is *not* being used to denote a deficit that is permanent or inevitable.) Structural deficits so defined were widespread in oil-importing developing countries following the two oil shocks—and if slow northern growth and high real interest rates are to be permanent phenomena, are now virtually universal in those countries. A structural deficit has a particularly strong claim to be financed with the aid of the international community while adjustment is effected. Given that the necessary adjustment program needs to combine a prudent demand management policy (and thus draw on the Fund's expertise) with structural adjustment (and thus draw on the World Bank's expertise), it seems logical that structural deficits be financed by parallel extended facility loans from the Fund and structural adjustment loans from the Bank. Complementary demand-side conditions would be negotiated by the Fund and supply-side conditions by the Bank.

2) The Fund has often been criticized for the distributional impact of austerity programs adopted under its guidance. It has tended to reply that it has no business interesting itself in the distributional consequences of members' programs, and (somewhat inconsistently) that in any event, the improvement in the rural-urban terms of trade that usually results from a more outward orientation and price liberalization is distributionally progressive rather than regressive. The second proposition is probably generally true, although there are countries where the major benefits of higher agricultural prices go to landlords rather than peasants and there are other aspects of Fund programs whose impact is typically regressive, so giving little reason to believe that Fund programs are typically progressive overall. The first proposition seems to me to rest on a confusion. No one is calling for the Fund to impose its views on income distribution on its members or to adopt the Gini coefficient as a new performance criterion. What is being suggested is that the Fund should build up, and subsequently offer to those members that wish to avail themselves of it, expertise in designing and implementing equity-oriented stabilization/adjust-

ment programs. One is happy to note that the World Bank does not suffer from the Fund's qualms on this issue (A. W. Clausen, 1982, p. 10).

3) The heart of the Fund's monitoring of the programs agreed with borrowing countries is through the performance criteria, any breach of which triggers a suspension of further disbursements. Performance criteria are normally framed in terms of domestic credit ceilings, as well as limits on credit to the public sector, limits on public sector deficits, and/or limits on public sector foreign borrowing. There are good reasons for the emphasis placed on such variables: they are policy variables, they can be objectively measured, the data to assess compliance are available promptly, a wide range of theories indicate that control of these variables is important to the payments outcome, and—in the case of the domestic credit ceiling—continued compliance with the ceiling when faced with an exogenous shock provides a stabilizing negative feedback to the balance of payments. Nevertheless, these appropriately conceived criteria appear to be applied by the Fund in an inappropriately rigid way. There are a wide range of unexpected events whose occurrence may make it appropriate to modify a target; for example, successful inflation stabilization will increase the demand for money and so justify a higher credit ceiling (especially in countries with low capital mobility). An even clearer example is provided by a country which obtains a large proportion of its tax revenue from export taxes and has accepted a ceiling to the public sector deficit as a performance criterion. An unexpected fall in export prices would then require the country to take further deflationary fiscal action to meet the target, which would in general be inappropriate; rather, the performance criterion should in that case be automatically modified. In general, the Fund should make a great effort, which it does not at the moment, to frame performance criteria as contingent conditions which will vary with the state of the world, rather than as fixed requirements.

4) Credit ceilings are a good technical way of checking compliance with overall

fulfillment of commitments to limit expenditure. The effectiveness of Fund monitoring would be increased if there were a similar assurance that the broad thrust of expenditure-switching policies would also be maintained through the program. This could be accomplished by adopting the real exchange rate as an additional performance criterion.

5) There have been marked variations in the toughness of IMF conditionality in recent years. As documented in my 1982 study, conditionality was eased in mid-1979 and tightened again in mid-1981. These variations in conditionality appear to have affected all its major dimensions: the length of programs, the severity of the retrenchment in expenditure that was required, and the prerequisite of a devaluation adequate to restore competitiveness. It is important to consider the merits of both the *timing* and the *content* of these variations in conditionality.

On the issue of timing, one can make a good case for the relaxation of mid-1979, since this coincided with the onset of the exogenously induced deficits created by the second oil shock and a new world recession, but no case whatsoever for the subsequent tightening of mid-1981, which occurred in the midst of deepening world recession. This judgment rests on the view that the Fund should seek to ameliorate the impact of world cyclical developments on its weaker members and in the process make a modest contribution to a global anticyclical policy—a view which appears to be regarded as heretical in the Fund. It is, however, a view which springs straight from the discussion of the rationale for Fund lending at the beginning of this paper, inasmuch as there is a lower world cost in easing external liquidity constraints at a time of world recession than at a time of boom.

On the question of content, the move to multiyear programs and the willingness to approve programs that did not involve the expectation of a short-run output loss were to be welcomed, and the retreat from those

positions in 1981 is to be deplored. In contrast, the weakening of the requirement for adequate expenditure-switching policies (in the form of the restoration of realistic exchange rates) in the period of easy conditionality was undesirable. The whole logic of Fund programs involves borrowing countries adopting measures that will suffice to restore a viable payments position in the medium term. There can be scope for legitimate differences of view about the state that the world economy should be assumed to reach in the medium term, but there can be no excuse for failing to insist on measures that give a reasonable prospect of adjustment being achieved even if the world economy is moderately prosperous. On the other hand, and of greater immediate relevance, the Fund will lose the ability to play any cyclical stabilization role if it takes too gloomy a view of the medium-term state of the world economy whenever the world develops a recession.

I conclude that there is no case for abandoning the basic structure of IMF conditionality, but ample scope for improving its application.

## REFERENCES

- Clausen, A. W., "Address" to the Board of Governors, Washington: World Bank, 1982.
- Dell, Sidney and Lawrence, Roger, *The Balance of Payments Adjustment Process in Developing Countries*, New York: Pergamon Press, 1980.
- Schmitt, Hans O., *Economic Stabilization and Growth in Portugal*, IMF Occasional Paper No. 2, Washington, 1981.
- Williamson, John, *The Lending Policies of the International Monetary Fund*, Washington: Institute for International Economics, 1982.
- \_\_\_\_\_, *IMF Conditionality*, Washington: Institute for International Economics, 1983.

# Devaluation: A Critical Appraisal of the IMF's Policy Prescriptions

By LOUKA T. KATSELI\*

It is by now well understood that the use of the extended facility of the Fund or approval for standby loans involve the undertaking of comprehensive programs of adjustment that "include policies...required to correct structural imbalances..." Even after the Fund's internal review of its guidelines on conditionality in March 1979, approval of standby agreements almost always requires severe tightening of expenditures through contractionary fiscal and monetary policy measures such as the imposition of ceilings on net government borrowing and/or net domestic assets of the central bank, or an upward adjustment of nominal interest rates in those cases where rates are fixed by the government. Approval also involves an "understanding" with the Fund on exchange rate policy and exchange rate arrangements. The understanding usually includes a substantial devaluation of the currency. Devaluation thus becomes part of a restrictive policy package which aims at improving the balance of payments of the country and its foreign exchange reserve position.

The advocacy of the same short-run demand-oriented policy package in different countries at different times has created the belief that underlying the policy prescriptions there exists a uniform IMF line of thought and, more importantly for our purposes here, a consistent and uniform analytical model. A cursory review of recent documents and mission reports, however, reveals that often the same policy prescriptions are based on different analytical arguments, some justified and some not, that cast some doubt on the widespread acceptance within the IMF of a particular theoretical structure as a guideline for policy. Thus by 1982, as theoretical advances in macroeconomics and in-

ternational finance have cast substantial doubt on the empirical if not theoretical validity of a simplistic monetary approach to the balance of payments, the analytical arguments that are used to sustain the traditional policy line have become by necessity murkier and more cumbersome.

This is nowhere more clear than in the analysis of exchange rate policy which is the focus of this short paper. The use of exchange rate adjustment as an active policy tool is critically analyzed in the following in terms of its effectiveness as a stabilization policy tool and as a substitute for tax or redistribution policy.

## I. Targets for Exchange Rate Policy and Macroeconomic Adjustment

One of the fundamental propositions of the macroeconomic literature pertaining to assignment of policy tools to targets of policy is that, in the absence of uncertainty, tools should be assigned to the target which they relatively affect the most. Thus for example, in the traditional analysis, monetary policy is assigned to the attainment of external balance and fiscal policy to that of internal balance.

The use and assignment of exchange rate policy revolves around the question of its relative effectiveness as a tool to improve the balance of payments as against a tool that simply translates the foreign exchange price of traded goods to domestic prices or vice versa. This realization serves to highlight the fundamental distinction between using the nominal exchange rate as an active instrument of external balance versus a tool that insulates the domestic economy from external goods-market disturbances. It is by now widely accepted that the assignment of the exchange rate to one of these targets depends on the structural characteristics of the econ-

\*Centre for Planning and Economic Research, Athens, Greece, and Yale University.

omy in question. Whether or not nominal exchange rate devaluation improves the current account depends on two channels through which it might have an effect: devaluation of the real exchange rate or a direct effect on domestic absorption.

The traditional Polak model (1957) and more recently the exposition in IMF (1977) that sketch the so-called IMF model depend on the latter channel. In this class of models, (a) all goods and assets are assumed to be perfect substitutes so that there is in essence one composite good whose price is given, (b) the country is a price taker in both goods and asset markets, and (c) all wages and prices are flexible so that output is fixed at its full-employment level.

Given these assumptions a devaluation works solely through a real-balance effect. The increase in the price of foreign exchange, reduces real balances. Since the demand for money (the only asset available) is exogenously determined, the excess demand for money is translated into a reduction in absorption and a balance of payments surplus which is the vehicle by which money balances are replenished. Thus in this model, "exchange-rate policy is really policy to manipulate the level of reserves, not the current account balance" (William Branson, 1982).

There is no room in this one-good model for changes in relative prices, improvement of trade competitiveness, or any related considerations. If anything, in this traditional monetary model the exchange rate affects the domestic price level while the level of real money balances affects the external balance.

Even if one stays within this broad framework, added considerations might mitigate the positive effects of the exchange rate change on the level of reserves at least in the short run. If prices are flexible and the economy very open, a devaluation can give rise to inflationary expectations that lead to an increase in consumption and imports especially of durables or an accumulation of inventories. Thus absorption might not be reduced as expected in the short run. More importantly in terms of its usefulness as a policy tool in the case of countries which have serious balance of payments problems, de-

valuation in terms of this analytical framework is at best a tool that validates past monetary expansion. If the nominal money stock could be restricted to its past level through tight policy, then devaluation would become unnecessary.

Finally, as noted by Branson, if a country faces exogenous fluctuations in interest rates or real output, monetary policy should be assigned to the maintenance of domestic price stability by offsetting the induced changes in the demand for real money balances. If disturbances originate in the foreign goods markets, then domestic price stability could be achieved through an off-setting move of the nominal exchange rate either via the market if the rate is floating or through direct policy intervention. Thus, in the context of the simple monetary version of the IMF model, the optimal exchange rate adjustment depends on the source of the disturbance both in terms of markets and location. This is of course a well-known result from the literature, but one which is often disregarded in actual policy prescriptions especially in the context of *LDCs* which face external disturbances in goods markets.

An altogether different channel by which exchange rate adjustment influences the current account is through its influence on a relative price often called the real exchange rate. The real exchange rate is equivalent either to the terms of trade in the context of a simple trade model with two countries that completely specialize in production, or to the relative price of traded to nontraded goods in the case of a small country two-good model. It has been shown that there are substantial differences both in the theoretical and empirical properties of the two indices (see my 1982b paper).

Increased competitiveness is associated either with a deterioration in the terms of trade which increases the export share of the country in the world market and reduces the quantity imported or with an increase in the price of the traded good sector. The story gets more complicated in the more realistic case of a country which is not a price taker in both or either markets but which also produces nontraded goods (see my 1982b paper).

In most IMF reports it is automatically assumed that a nominal devaluation will increase competitiveness by affecting the real exchange rate. It is evident that substantial effort has been recently devoted to proper measurements of the real exchange rate especially in countries with diversified trade leading to the presentation of alternative indices of real overvaluation of the currency. The appropriate choice of base period, of weights (import, export, total trade or some combination of *MLRM* weights) and of the domestic and foreign price index to be used (wholesale price indices, unit labor costs, etc.) become the object of much discussion and guarded analysis in appendices of mission reports. The choice of the relevant countries also becomes critical with a growing preference towards inclusion of countries which compete with the host country in third markets rather than of the bilateral trading partners. Both the choice of countries and weights depends on the question that is being posed.

Technical questions apart, what is rarely seriously discussed in negotiations or in written reports is the potential effectiveness of the nominal exchange rate on the real exchange rate. This is so at a time when the theoretical literature abounds with examples of cases where the structure of the economy is such that the real exchange rate is not affected by nominal exchange-rate movements and a devaluation is thought to have negative effects on both output and the price level.

Two extreme examples might be sufficient to demonstrate how the structure of the economy might be such that a nominal devaluation will have no effects on the real rate. Suppose that we consider a less developed country that imports intermediate goods to be used in domestic production under fixed coefficients and exports agricultural goods or raw materials whose supply is inelastic in the short run. This simplification might be pertinent for a country like Madagascar, the Sudan or even Kenya, all of which have negotiated standby agreements with the IMF. Thus in this case both the supply elasticity for exports and demand elasticity for imports approach zero. A deval-

uation will not affect the terms of trade and the effect on the balance of trade will depend on the initial trade balance. Given a large initial deficit, the exchange rate adjustment will magnify it. The quantity exported might even be reduced if intermediate imports are used in the export sector leading to even more perverse effects. The increase in the cost of production of nontraded goods might also prevent the expected increase in the relative price of traded to nontraded goods and lead to internal stagflation depending on the relative elasticities (see my 1982a article). Thus in the presence of intermediate imports the argument that devaluation is a useful tool for promoting competitiveness becomes at best uncertain as profitability is impaired through the increase in the cost of production while the terms of trade are not seriously affected. These considerations are especially important for some of the smaller *LDCs* which do not have large import competing sectors and have structures of trade which are not much different from those in the example above.

Even in the absence of intermediate goods, full nominal wage indexation is sufficient to prevent a change in the relative price of traded to nontraded goods. If nominal wages are tied to the consumer price index, then a devaluation will increase nominal wages and the price of nontraded goods by the full amount of the devaluation. The recent emphasis on the supply side has thus shifted the focus of discussion of the implications of exchange rate adjustment from demand towards the cost of production. The shift in emphasis has prompted a serious questioning of the effectiveness of nominal exchange rate adjustment on the current account and its role in macroeconomic adjustment. This theoretical questioning is still open to empirical testing, but it is fair to say that its powerful and controversial messages have not yet been adequately incorporated into IMF thinking.

While the role of the exchange rate in macroadjustment is debated, three additional roles for exchange rate policy have emerged. These include its substitute role as tax policy especially in the context of *LDCs*, its role in income redistribution, and finally its role in investment promotion and development in general.

## II. Alternative Roles for Exchange Rate Policy

For a nominal devaluation to improve the current account through reduction in absorption and/or a change in relative prices, the domestic price of traded goods and the general price level must increase. If domestic prices are fixed by the government this channel of adjustment is blocked. In that case, as the government sells goods at a world market price which is higher than the producer price it pays, exchange rate policy becomes a substitute for tax policy. A devaluation of the currency simply increases export proceeds and serves the function of an export tax. Alternatively, a policy to raise domestic producer prices to the world market level and to devalue by the same amount increases the export tax in absolute value (depending on the supply elasticity) and provides incentives for export production. This was in fact the analytical basis of the IMF's proposal for devaluation in the context of negotiations with some Eastern African countries and this reasoning once again had little to do with competitiveness considerations. The objective here was to increase the quantity exported but maintain budget revenues as far as possible. This use of exchange rate policy as tax policy becomes especially relevant for countries which export few agricultural commodities or raw materials that are traded by the central government and which lack the tax base for an effective income tax policy. Devaluation is thus linked to price liberalization and efforts to mitigate the unwarranted effects resulting from the correction of domestic relative price distortions. There are problems however with this approach as well.

Given the fact that devaluation affects import prices and specifically the prices of imported inputs, the effectiveness of the policy package critically depends on the profitability of the export sector. As discussed earlier, it is the effective protection rather than nominal protection that matters and taxation of inputs reduces profitability; so does the increase in labor costs that arises in an indexed economy.

It should be pointed out that a devaluation also increases the home currency value of interest payments on public external debt,

making unclear the net effect on government expenditures. More importantly, the existence of a marketing board usually gives a country some market power that is not attainable if trade is undertaken by many small competitive producers. If that is the case, then the country is "semi-small" and a devaluation that is proportional to the increase in domestic producer prices will deteriorate the terms of trade and might lead to a possible reduction in tax revenue and net export receipts.

Exchange rate devaluation has also been thought of as a stimulus to private saving due to the expected redistributive effect from wages to profits, that is, to a sector which is assumed to be characterized by a higher marginal propensity to save. This line of argument which is implicit in some of the earlier IMF documents has been criticized on many different grounds. Putting aside the fact that the marginal propensity to save among profit earners is not necessarily higher than among wage earners, an increase in the profitability of the traded good sector critically depends on the ensuing increase in the cost of raw materials and labor as well as on the country's relative degree of market power.

In terms of the earlier discussion on macroeconomic adjustment, a necessary condition for redistribution is that the relative price of traded to nontraded goods changes. In that case, a devaluation will benefit those factors of production that are intensively used in the production of traded goods and the consumers of nontraded goods. In the case of a country which produces and exports primary products as is the case of many *LDCs*, the main beneficiaries of the devaluation will probably be the export crop growers insofar as domestic export prices are allowed to increase.

These considerations highlight the role of exchange rate policy in the context of an overall development strategy. It can be argued that if the objective is to promote investment and industrialization the possible short-run benefits on the balance of payments brought about by exchange rate devaluation should be weighed against possible longer-run costs. Given the importance of imported capital goods at initial stages of

development there might be a role for maintaining a slightly overvalued currency for some periods of time, especially if the internal tax and transfer system is not adequately developed. This strategy was adopted at some early phases in the development of Japan and some of the newly industrialized countries, and its merits should be judged in light of the structural characteristics of the economy in question.

More importantly, stabilization policy should not be viewed as a substitute to development policy. It is often the case that constraints in development, such as foreign exchange availability or insufficient domestic saving, are perceived by international credit organizations as the targets of policy. It is not clear why a developing country should thrive to reduce the deficit in its current account or what the criteria should be for doing so. It is clear, however, that there should be a long-run steady-state path that policy should be aiming for depending on a country's level of development, but such considerations have not yet been seriously addressed by the IMF. In a development context, the focus should shift from the current account to the basic balance where the required net long-term capital inflow for development requires a current account deficit that is supported by an appropriate real exchange rate. In that framework, the equilibrium real exchange rate is determined by long-run growth and investment prospects

and nominal exchange rate policy is adjusted accordingly. It is thus imperative not only to reconsider the short-run effectiveness of macro-adjustment policy in light of new theoretical and empirical knowledge, but also to place stabilization policy in a dynamic context.

## REFERENCES

- Branson, William H., "Economic Structure and Policy for External Balance," in *International Monetary Fund Conference on Policy Interdependence*, National Bureau of Economic Research, August 1982.
- Katseli, Louka T., (1982a) "Macroeconomic Adjustment and Exchange-Rate Policy in Middle-Income Countries: Greece, Portugal, and Spain in the 1970's," in M. de Cecco, ed., *International Adjustment, the EMS and Small European Countries*, Blackwell, October 1982.
- , (1982b) "Real Exchange Rates in the 1970's," Discussion Paper No. 403, Economic Growth Center, Yale University, May 1982.
- Polak, Jacques, J., "Monetary Analysis of Income Formation and Payments Problems," *IMF Staff Papers*, November 1957, 6, 1-50.
- International Monetary Fund, *The Monetary Approach to the Balance of Payments*, Washington: IMF, 1977.

## SPECIAL REPORT ON BOOK PUBLICATION

### On Contracting with Publishers: Author's Information Updated

By MARTIN SHUBIK, PEGGY HEIM, AND WILLIAM J. BAUMOL\*

In the AAUP *Bulletin*<sup>1</sup>, March 1967, William Baumol and Peggy Heim presented the results of a survey in which over 350 faculty members were polled to determine the types of books they had published, the royalty rates they had received, the payments made to them for reviewing or editing of manuscripts, the general relationships between themselves and the publisher. These results gave a reasonably clear picture of the circumstances of the author.

It was learned that the academic author is generally well satisfied with what the publisher has done for him—to an extent perhaps even surprising in a profession not noted for the reticence of its criticism. Nevertheless, as was to be expected, a number of unfortunate cases, some of them involving rather questionable practices, did turn up. Several such cases were described in that article and more are noted here.

In this report, we update and, to some extent, specialize the previous study using as our data a questionnaire that we designed and circulated to the chairmen of 116 economics departments and the deans of 62 business schools asking them to circulate the questionnaire to a random sample of the members of their department or school. Replies were received from 94 authors at 44 economics departments and 71 authors at 23 business schools, representing 38 percent of the departments and 37 percent of the schools contacted.

\*Cowles Foundation, School of Organization and Management and Department of Economics, Yale University; TIAA-CREF; Princeton and New York Universities; respectively. This report was prepared at the request of the Executive Committee of the American Economic Association. We are grateful to Gary Burke, Harcourt Brace Jovanovich, and Leo Raskind, University of Minnesota Law School, for their comments.

<sup>1</sup>We wish to thank the AAUP for permission to incorporate substantial parts of the text of the *Bulletin* article, "On Contracting with Publishers: Or, What Every Author Should Know," vol. 53, into this report.

Although publication of books is an important part of academic life, information on the terms on which publication can be arranged is surprisingly scarce.<sup>2</sup> For example: few authors seem to be aware of the fundamental fact that—particularly where a contract is signed before the manuscript has been completed—usually, and almost unavoidably, the publisher is not committed to bring out the book in question, though the author is committed to offer his manuscript to the publisher with whom the contract is signed; that the contract binds the author without really committing the publisher.

Following the format of the Baumol-Heim article, this report undertakes to supply some of the pertinent information in a more or less systematic manner. We have attempted no exhaustive analysis of all the formalities of the contract, nor have we pursued in depth the legal status of copyrights or any similar matters. Rather we have sought to collect information which most faculty members are likely to find useful in arriving at an arrangement with their publishers. We describe the range of those financial terms that seem to be of most widespread interest—royalties (including the rather confusing and not read-

<sup>2</sup>J. C. Hogan and S. Cohen, *An Author's Guide to Scholarly Publishing and the Law*, Englewood Cliffs: Prentice-Hall, 1965, especially ch. 4; and O. Cargill, W. Charvat, and D. D. Walsh, "The Publication of Academic Writing," *Publications of the Modern Language Association of America*, September 1966, 81, pp. 39–45. Cargill et al. is primarily a brief compendium of items of good advice to the young author. For a legal treatise on copyright, see Melville B. Nimmer, *Nimmer on Copyrights*, New York: Matthew Bender, & Co., Inc., 1981. See also Paul W. Kingston and Jonathan R. Cole with Robert K. Merton, *The Columbia University Economic Survey of American Authors: A Report of Findings*, New York: Columbia University Press, 1981, for economic status of writers; and for general information for any author on how to deal with publishers, see Judith Appelbaum and Nancy Evans, *How to Get Happily Published*, New York: Harper and Row, 1978.

## PROPOSED INFORMATION FORM FOR AUTHORS

from Period \_\_\_\_\_ to \_\_\_\_\_

Name of Book: \_\_\_\_\_

Name of Author: \_\_\_\_\_

Name of Publisher: \_\_\_\_\_

1.

Sales	Number of Copies (1)	List Price (2)	Publishers' Proceeds* (3)	Royalty Terms (as ____% of which col.) (4)	Royalty (5)
Domestic:					
Trade					
Text					
Foreign					
Other types					
Total sales					

Cumulative total sales from date of publication:

Of current edition \_\_\_\_\_

Of earlier editions \_\_\_\_\_

2. Other types of proceeds to publisher

(1)

Payment to Author

(2)

(1) Rental of plates, sales of sheets, etc.

(2) Translation or other editions (including paperbacks) already contracted for

(3) Proceeds from licenses, fees, etc. not covered above

Total of "other proceeds"

3. Translations or other editions contracted for:

Language or Other Specification (1)	Publisher (2)	Payment to Author (3)
(1) _____	_____	_____
(2) _____	_____	_____
(3) _____	_____	_____
(4) _____	_____	_____

4. Number of books in stock \_\_\_\_\_ as of (date) \_\_\_\_\_ \*\*

\*To be supplied only where royalties are based on proceeds.

\*\*The author will be notified if it has been decided to let the book go out of stock, and will also be notified when it is decided to reprint so that he can submit lists of any errors for correction.

FIGURE 1

ily comparable variety of bases on which they are calculated), advances, cost of corrections, and discounts on copies of the book purchased by the author.

We reiterate the two proposals made some fifteen years ago to publishers. We believe that they are still germane and can contribute substantially to better relations between publishers and their authors.

We suggest that publishers provide better information on the base in terms of which royalties are calculated. Furthermore, we suggest the adoption of a standardized information form (Figure 1) which will report to the author data that are highly important to him, but which are not now supplied to him as a matter of course by all publishers. We believe such a form will impose no serious burden, will help to avoid needless misunderstandings, and enable both publishers and authors to perform their work more effectively.

Unfortunately, there are some problems of publication that cannot be handled simply by contractual provisions. Neither an information form nor a contract can guarantee that copy editing will be of satisfactory quality or that promotion will be vigorous or that the printer's work will be good. Publishers vary enormously in the effort they devote to the quality of the volume and to its sale, and an author may often find it preferable to select a house whose record on these matters is very good, even if its royalty rates are lower than those offered elsewhere. On all these issues there seems to be no substitute for checking past performance. Before signing with any publisher, an author is therefore well advised to speak, if he can, with one or more people who have had books published by the firm in question, and to examine other more or less similar volumes that the publisher has put out, as well as some of the company's relevant advertising and other promotional material. There is probably no more reliable way than this to obtain information on the quality of a publisher's service in its many aspects—concerns of great significance for the author which are difficult, if not impossible, to cover explicitly by contractual provisions.

### The Survey and Its Interpretation

As a way of enlarging upon the reasons for our recommendations, the results of the survey are presented. We note features common to all authors, yet contrast economics departments and business schools where relevant.

*Who Are the Publishers?* Table 1 reports the leading publishers of the books written by our respondents (entries in parentheses indicate number of responses). There are no surprises here. For 179 books published by members of economics departments, 66 presses were used. The 128 books published by business school faculty who replied were published by 55 different publishing companies.

*Subject Matter and Type of Book.* Table 2 reports the subject matter of the books.

TABLE 1

Economics Departments	Business Schools
A. Top Five University Presses	
Johns Hopkins (10)	Johns Hopkins (1)
Ohio State (3)	NYU (1)
Oxford (3)	University of Nebraska (1)
University of Texas (3)	No others listed
University of Michigan (2)	
Harvard (2)	
MIT (2)	
B. Top Five Other Presses	
McGraw-Hill (12)	Prentice-Hall (11)
Wiley (10)	McGraw-Hill (11)
Heath/Lexington (10)	Irwin (10)
Ballinger (9)	Wiley (10)
Prentice-Hall (9)	Dryden (4)
	Houghton Mifflin (4)

TABLE 2

Economics Departments	Business Schools
Top Five Categories	
Economics (132)	Economics (31)
Economic History (7)	Finance (7)
Finance (5)	Marketing (7)
Statistics (5)	Management (7)
Economics/Law (3)	Accounting (7)
Accounting (3)	
Management (3)	

TABLE 3

	Economics Departments	Business Schools
Freshman text	11	9
Advanced undergraduate text	54	53
Graduate text	38	42
Popularized book in own field	11	12
Technical or specialized books for professional audience	84	39
Collection of readings edited by you	17	7
Papers prepared for conference, edited by you	13	6
Other	10	3

They fall into 26 and 49 subfields for economics departments and business schools, respectively. These included topics such as (general) economic issues, sales management, women in business, labor unions, public finance, economic development, antitrust economics, and economics of resources.

The information in Table 3 shows that there was a greater emphasis upon technical and research books at economics departments and a somewhat higher percentage of advanced undergraduate and graduate textbooks at schools of business.

*Manuscript Submission.* A problem faced by many authors involves time delays between completion of a manuscript and its final acceptance for publication. Somewhat over 25 percent of the authors have dealt with this problem in part by sending their manuscripts to more than one publisher simultaneously. (See Table 4.)

We believe that, given the probability of delays, multiple submissions are ethical if disclosed. It has been suggested by Leo Raskind, counsel for the AEA, that a sentence of the following variety in a letter accompanying a submission is appropriate. "In submitting this manuscript to you, I have not foreclosed the possibility that I might submit this for consideration to another publisher during the course of your deliberations. I trust this is acceptable to you."

*Acceptance and Publication Time.* Tables 5 and 6 illustrate the time lags encountered between submission and acceptance, and

TABLE 4

	Economics Departments	Business Schools
Submitted to one publisher at a time	132	90
Submitted to more than one publisher at a time	46	35

TABLE 5—SUBMISSION OF MANUSCRIPT TO ACCEPTANCE

	Economics Departments	Business Schools
Under 1 month (or in advance)	8	19
1 month	15	17
2 months	15	10
3 months	29	19
4 months	14	5
6 months	28	6
9 months	4	2
12 months	9	1
Longer	6	—
No reply	45	36

TABLE 6—TIME FROM ACCEPTANCE UNTIL PUBLICATION

	Economics Departments	Business Schools
2 months	1	—
3 months	5	1
4 months	—	12
6 months	23	18
9 months	22	34
12 months	38	5
15 months	13	6
18 months	5	8
24 months	9	1
30 months or more	3	—

acceptance and publication. The replies indicate that the speed of acceptance (or rejection) is far greater than is suggested by folklore.

The time between acceptance and publication has a clear mode at one year, but as can be seen, there are considerable "tails" to the distributions beyond one year and it appears that some authors have had difficulties as is indicated by the following comments.

Very serious delay occurred several years ago with a manuscript submitted to the University of \_\_\_\_\_ Press. The Press held the MS for over a year, without an answer; I finally insisted they return the MS to me and it was published by another firm. I had had very good relations and experience with Irwin, Johns Hopkins, MIT, and Stanford Presses.

On a (yet unpublished) book we had a signed and sealed contract with \_\_\_\_\_ (a commercial press). When it came time to publish, ...informed us that because of *their* difficulties, they were unable to go ahead with publication. To this day, we have been unable to obtain any kind of satisfaction.

Most recent horror stories which have come to my attention entail mergers in the industry. E.g., an author signs a contract and completes a substantial portion of a manuscript only to be told that "his" company has been merged and the new parent firm is no longer interested in the manuscript. I presume the author has legal recourse, but that can be costly and of uncertain outcome. I suspect the moral here is that it is not a bad idea to take a sizable advance so the publisher has a substantial stake in the project being completed and the book ultimately published.

It is suggested that in a contract a sentence or two of the following sort may help:

The publisher agrees to use best efforts to produce this manuscript timely and expeditiously in order that the content of this manuscript not be outdated on publication. In the event of a delay of more than \_\_\_\_\_ days from the date of acceptance of the manuscript to the beginning of production and manufacture, the publisher agrees to return this manuscript to the author on request.

*Type of Publication.* As can be seen from Table 7, most of the publications are still in hardcover.

TABLE 7—TYPE OF PUBLICATION

	Economics Departments	Business Schools
Hardcover	70%	75%
Paperback	16	14
Both	14	11

TABLE 8—TIME OF PUBLICATION

	Economics Departments	Business Schools
Last five years	57%	63%
Previous five years	28	28
Earlier	15	9

TABLE 9—SIZE OF SALES

	Economics Departments	Business Schools
Not yet reported	4%	6%
1-2,000	46	25
2,001-5,000	22	24
5,001-20,000	20	29
Over 20,000	8	16

*Size of Sales.* The books reported upon had, for the most part, been published in the last five years. Table 8 indicates the breakdown.

The sales figures in Table 9 indicate that most of the scholarly and research books are destined to have relatively modest sales in the range not exceeding 2,000. The business schools offer more market opportunities than do the economics departments as is indicated by the 16 percent of books published whose sales exceeded 20,000.

#### Contracting Conditions

Of prime concern to the author is the actual contract obtained. This includes not only the royalty agreement, but a host of special conditions.

*Types of Financial Arrangements.* Standard financial arrangements take four main

forms: (1) royalty agreements; (2) sale-of-rights agreements; (3) stock-option agreements; and (4) subsidy publishing agreements. The majority of the books reported upon here involved royalty agreements.

*The Royalty Agreement.* Basically, this agreement, which is the most common arrangement in use today, provides for the payment of some specified percentage of list price, gross proceeds, or net proceeds on the number of copies sold. Generally, terms are not the same for domestic and foreign sales, and the contract spells out the degree to which the author shares in the proceeds from the sale of subsidiary rights. Under most circumstances, the author will probably find the royalty agreement more advantageous than any of the other three forms of compensation agreement described here. (There are exceptions, however; for example, a book with wide appeal which is marketed effectively under a private publication agreement.)

*The Sale-of-Rights Agreement.* Under this agreement, the author receives a fixed sum for which he assigns to the publisher the copyright and all other rights and interests in the book. In effect, the author receives a flat fee for his writing service and the use of his name. Disposition of the product is then in the hands of the publisher. The sale-of-rights contract generally offers the author very little protection. Unless the contract explicitly provides a procedure for revised editions, the publisher can appoint a second party to revise the book (generally a text, an industrial manual, etc.), and the original author, without resorting to potentially costly litigation, and perhaps not even then, has no more control over the revisions than the publisher is willing to grant him. Indeed, the problems to which a sale-of-rights contract may lead can be so troublesome that the author is well advised to consult legal counsel before signing any such agreement.

*Stock-Option Agreements.* These are usually used only by small or new publishing houses. The agreements are essentially profit-sharing agreements in which the author receives compensation in the form of stock shares. The value, of course, depends upon the success of the company and upon the existence of a market for the stock. The

author may also receive stock options for the purchase of additional shares. In some cases, royalty agreements are combined with some form of stock arrangement.

*Publishing Subsidy Agreements.* Many contracts call for direct payments to the publisher by the author before the book is brought out. In our survey, some 10 percent of the contracts required some subsidy payment ranging from \$200 to \$8,350, with an average of \$2,090 for hardcover books.

These subsidies are designed to permit publication of manuscripts which might otherwise not be financially feasible for the publisher to undertake, or, if the author expects very large sales, to insure that the bulk of the profits go to the author, or to make the book available at a price lower than the one at which it would otherwise have been sold. While often such subsidy agreements are entered into with standard publishing houses, particularly with university presses, they are most frequently employed by what is known in the trade as the "vanity press," which handles purely private publication which authors arrange entirely on their own behalf.<sup>3</sup>

According to our survey, the most common complaint is that the net cost to the author was underestimated, generally because sales, and therefore royalty income, were overestimated, despite relatively high royalty rates designed to make the contract look attractive. In another instance, the agreement obligated the author to pay the costs without stipulating a maximum payment by the author, and the cost of publication exceeded the estimate by a significant amount.

In publication by standard presses, direct prepublication payments to publishers are sometimes requested to permit the book to

<sup>3</sup>If a manuscript is not of interest to a commercial publisher or a university press but the scholar nevertheless wants to arrange for its publication, he may find it useful to consult *Literary Marketplace: The Business Directory of American Book Publishing*, which provides a list of commercial printers who perform this service and the names of independent agents, who, at author's expense, give technical advice (for example, on book design, pricing policy, etc.), and who handle promotion and advertising.

be sold at a price lower than would otherwise prevail. However, a number of authors who provided prepayments on these grounds indicated that their lower price expectations were, in fact, not realized. Clearly this may represent a misunderstanding and failure of communication on the part of both parties. However, if a lower selling price is the object, it would appear to be desirable to have some specific price agreement in writing from the publisher rather than to rely upon some general reference to "keeping the price low," which may mean different things to author and publisher.

On occasion, the special research funds available to scholars in some of the more progressive and higher-paying liberal arts colleges and private universities are also used to help subsidize publication, particularly where the institution does not have its own press. Sometimes, when the subsidy is paid by an organization or institution, a special provision may be included in the contract stipulating that payment of royalties to the scholar will not begin until the subsidy provided by the institution or by a foundation has been repaid.

#### **Royalty Rates: Bases Used in Calculating Basic Royalty Payments**

Royalty payments are usually calculated as a percentage of one of three alternative bases—list or retail price, gross proceeds (sometimes called net price), and net proceeds. The term "gross proceeds" generally refers to list price less discounts to book-sellers.<sup>4</sup> While the difference between gross proceeds and net proceeds is not clearly defined, net proceeds, as used by some publishers, means gross proceeds less other discounts, returns, and allowances—items of significant magnitude in some instances. Apparently, other publishers consider the two terms synonymous.

<sup>4</sup>Where books are sold by the publisher directly to the final purchaser at list price (as when books are sold by mail), the corresponding portion of gross proceeds is identical with list price. Often a substantial proportion of the sales of a volume is of this variety.

*Royalty Bases and Rate Structures.* Because publishers use different systems of discounts and pricing, depending upon whether a book is sold as a trade edition or as a text—there are other variations too which we shall not attempt to explain here—there is no easy and accurate way to describe royalty rate structures for purposes of comparison. The system places the author at a great disadvantage in his effort to evaluate the royalty terms proposed in his contract.<sup>5</sup> Thus the apparently simple conversion of list-price-based royalties on a college text to the equivalent royalty rates on gross sales normally runs into a major complication. In addition to sale as an adopted text, the book (usually with a different cover and jacket) is often sold in a trade edition at a higher list price—about one-third more than the text—and at a higher discount to the bookstore. This discount to the bookstore commonly runs about 20 percent of the textbook's list price for the educational edition and about 40 percent of trade list for the trade edition.<sup>6</sup> Yet the two editions are commonly priced to yield approximately the same gross proceeds to the publisher. As a result, an author whose royalties are based on gross receipts will do equally well from either type of sale, but an author paid a fixed percentage of list price will earn more from the sale of a trade edition copy than from a copy of the text version if the royalty rates on both editions are the same. Of course, if, as publishers suggest, sales of textbooks in trade editions are often relatively small and hence unprofitable, the author may well find it desirable to accept a lower royalty rate on the trade volume in order to induce the publisher to provide such an edition.

<sup>5</sup>The author can obtain some useful information by asking the publisher with whom he is negotiating for sample information on royalty rates and total royalty payments on a volume similar to his own.

<sup>6</sup>A further complication is introduced by the "agency plan" which is apparently becoming more important for academic authors. A bookseller who undertakes such an arrangement is given a 33 $\frac{1}{3}$  percent discount on books that would otherwise be sold to him at a "short," say 20 percent, discount, but in return he commits himself to handle some minimum number of copies of every book the publisher produces in the given subject area.

TABLE 10—ILLUSTRATIVE CONVERSION TABLE FOR GROSS (OR NET) PROCEEDS: ROYALTY RATE EQUIVALENTS

Average Gross Proceeds as Percent of List Price	Average Gross Proceeds Equivalents for List Price Royalty Rates		
	10% of List	12% of List	15% of List
80 percent	12.5% of Gross	15.0% of Gross	18.8% of Gross
75 percent	13.3% of Gross	16.0% of Gross	20.0% of Gross
70 percent	14.3% of Gross	17.2% of Gross	21.5% of Gross
60 percent	16.7% of Gross	20.0% of Gross	25.1% of Gross

*Royalties on Specialized, Technical, and General Books.* Scholarly volumes which are not textbooks are priced for sale to the trade (i.e., for sale by regular retailers). Hardcover volumes frequently, but not always, are sold to retailers at a price equal to  $66\frac{2}{3}$  percent of list price (a discount to the bookstore of about  $33\frac{1}{3}$  percent). The discount is usually less than 40 percent, and sometimes goes as low as 20–25 percent on single copy sales.

A conversion table may help the scholar compare royalty rates based upon list price with those quoted on another base, such as gross proceeds, when the latter is a more or less fixed proportion of list price (for example, where gross proceeds are 80 percent of retail). Of course, because of the complex system of pricing and discounts any such calculation must be an oversimplification which does not take into account retail price differences between books sold to the trade primarily as technical, specialized, or general works and those sold as college texts.

Table 10 indicates, for example, that if gross proceeds are 70 percent of retail price, the author would need royalties of 21.5 percent of gross proceeds in order to be as well off as he is with rates calculated at 15 percent of list price.<sup>7</sup>

*Fixed or Graduated Rates.* The basic rate may be stated as a fixed percentage on all copies sold, or the rate may be graduated upwards as sales increase, for example, it may be set at 10 percent of the first 5,000 copies, 12.5 percent of the next 5,000, and 15 percent of the remainder. There are various

reasons for the use of graduated rates. If the book is going to be published anyway, the procedure may serve to reduce the financial risk to the publisher; if publication is uncertain, the arrangement may help a specialized book with limited sales appeal see the light of day. On the other hand, one can look at the provision through publishers' eyes and say that it gives the author an opportunity to participate in the profits derived from a larger volume of sales.

The graduated rates usually employ no more than three steps, but the number of thousands of copies involved varies significantly among publishers and even among contracts with the same publisher. Most commonly, royalties begin with the first copy sold, but a wide variety of alternative arrangements exist. Royalties may be omitted on the first 1,000 to 3,000 copies sold, or, on occasion, this number may be extended to 5,000 or even 7,000. Some contracts go a step further and provide that no royalty will be paid until the publisher recovers the cost of producing the book. Under the latter provision, the scholar relies on the good faith of the publisher, generally a university press.

Another variation on the graduated rates uses a less favorable royalty base, as well as a lower rate, on the first several thousand copies sold. Simply by changing an adjective the effective basic rate can be lowered significantly without its being very apparent to the author. Table 11 illustrates the real effect of the arrangement.

This technique is, of course, used to merchandise the royalty package. By changing the descriptive term rather than the figure, the rate structure can be made to look more attractive than it really is. Apparently,

<sup>7</sup>Some contracts utilizing gross proceeds as the base state that the term shall be construed to mean some specific proportion, such as 75 percent of the list price.

TABLE 11

		List Price Equivalent
First 3,000 copies	10% of net price <sup>a</sup>	7% rate
Second 3,000 copies	10% of list price	10% rate
Above 6,000 copies	15% of list price	15% rate

<sup>a</sup>Defined as 70% of list.

this practice is relatively new and uncommon.

It may be noted here that the direct financial return from the publication of a volume is generally not the most important consideration to the scholar, and this may help to account for the fact that many of the respondents in our survey did not indicate what royalty base is used in the calculation of their royalties. In the opinion of some authorities, publication of a first book may provide returns in the form of higher academic salary far greater than any royalty payment. It has been surmised that having a widely distributed book to his credit may well be worth many thousands of dollars to the young faculty member in terms of increased lifetime earning power, aside from any royalty payments.

In any event, the preceding discussion suggests very clearly that a high confusion cost is incurred by the use of three different bases for the calculation of royalties—in terms of retail price, gross, and publisher's net. Whether the result is deliberate or unintentional, it is obviously unfortunate and constitutes a major obstacle to the author's deliberations.<sup>8</sup>

<sup>8</sup>We are told that textbook publishers are now considering a proposal under which all books would be sold to bookstores at a fixed net price on which all royalties would be calculated. The stores would then make the retail price decision entirely on their own. Several publishers have already adopted this at the request of the bookstores.

We observe here that adoption of this practice by book publishers customarily using list price in the calculation of royalties can have important implications for their authors. Our data suggest that when gross and net

It is surely desirable for publishers to provide sufficient information to authors to permit them to determine the composition of net and gross proceeds as these terms (or their counterparts) are used by the firm. Only then will an author be in a position to evaluate competing proposals from two publishers. At a minimum the author should be supplied with information on the average relationship of retail price to gross and to net, so that he can perform a translation from one to the other for himself. Surely an author will have some justification for the suspicion that any publisher who refuses to do so must have something to hide.

*Royalty Payments in Practice.* Which royalty base is used most frequently in the calculation of royalties is apparently influenced by how one makes the count. In their earlier study, Baumol and Heim reviewed standard contract forms for seventeen relatively well-known commercial publishers and, by coincidence, for seventeen university presses. They found that about 60 percent of the publishers used list price as the royalty base for hardcover books, another 6 percent used two different bases, one of which was list price, and the remaining 34 percent used gross or net proceeds. The percentage breakdown by type of publisher is shown in Table 12.

In the current survey, we focused not on the publishers and their standard contract forms, but on books written by the survey respondents in economics departments and schools of business. As shown in Table 13, gross or net proceeds were used in royalty calculations for almost three-fifths of the books and list price in the others.

There was little difference in royalty base among the three major book categories—undergraduate texts, graduate texts, and specialized or technical books for a professional audience. The royalties of the majority of

proceeds were used in the calculation of royalties, rates at the upper end of the rate spectrum were not adjusted sufficiently on many books to compensate for the choice of the lower base. (See Table 16 and related discussion.) Only time will tell whether a similar problem will arise here.

TABLE 12

Royalty Base Used by Publisher	Type of Press		
	Commercial	University	Combined
List price	53%	67%	60%
Gross or net proceeds	35	33	34
Both list price and gross or net proceeds	12	0	6

TABLE 13

Royalty Base Used by Publisher	Type of Press		
	Commercial	University	Combined
List price	42%	42%	42%
Gross or net proceeds	58	58	58
Number in sample	203	26	229
Percentage distribution among presses	89%	11%	100%

books were based on gross or net proceeds, and list price was used in less than two-fifths of the books in each category.

In negotiating a contract with a publisher, authors often want to know how the royalty rates on their books compare with rates received by others. Since any one author may have different rates on different contracts, we have analyzed these on a book-by-book basis for the 178 books with complete data. The findings are summarized in Table 14, which shows the variation in royalty rates at two specific sales levels—the 5,000th and 10,000th copy—by type of book and the base used in calculating royalties.

If one looks at the minimum and maximum royalty rate percentages at 5,000 copies, the lowest percentage turned up in the survey was 3.5 percent of the list price for the editing of a collection of readings; the highest, 20 percent of net proceeds for a combination undergraduate/graduate text. Our survey turned up no royalty rate figure higher than 20 percent, whatever the volume of sales. Of course, this does not mean that books in economics and business never carry rates higher than 20 percent.

We note here that some books had multiple uses. Since we could not distinguish between primary use and distinctly secondary

use—for example, when a specialized book is also used as a collateral text in some graduate courses—the rates were included in each classification. This complication may have muddled the statistics, and we suggest that in any future survey, respondents be asked to distinguish primary from secondary uses.

A question that often arises is whether royalty rates are higher for one type of book than another. As Table 14 shows, the arithmetic mean royalty rate for any given royalty base was usually higher for undergraduate texts than graduate texts and both were higher than those for technical or specialized books. The difference between undergraduate and graduate textbook means were small and not statistically significant. When technical or specialized books were compared separately with undergraduate and with graduate texts, using the *t* test, the difference between means was generally significant (at better than the 1 percent level). The exception was with the specialized-book/graduate-text comparison, using a list price base (results mixed—significant at the 3 and 10 percent levels).

For textbooks—both graduate and undergraduate—the royalty rate at 5,000 copies was almost never less than 10 percent. In-

TABLE 14—ROYALTY RATES BASED ON LIST PRICE, GROSS PROCEEDS, AND NET PROCEEDS, FOR THE 5,000TH AND 10,000TH COPY SOLD, BY TYPE OF BOOK

Type of Book and Statistical Measure	Rate on List Price		Rate on Gross Proceeds		Rate on Net Proceeds	
	5,000 copies	10,000 copies	5,000 copies	10,000 copies	5,000 copies	10,000 copies
Undergraduate Text						
Mean	14.6%	14.7%	14.6%	15.1%	15.3%	15.7%
Median	15.0	15.0	15.0	15.0	15.0	15.0
Minimum	10.0	10.0	10.0	10.0	10.0	10.0
Maximum	18.8	18.8	18.0	19.0	20.0	20.0
No. in Sample	31	31	14	14	40	40
Graduate Text						
Mean	13.9	13.9	15.5	15.9	14.8	15.4
Median	15.0	15.0	15.0	15.0	15.0	15.0
Minimum	6.0	8.0	12.0	12.0	10.0	10.0
Maximum	18.0	18.0	18.0	18.0	20.0	20.0
No. in Sample	23	23	9	9	30	30
Technical or Specialized Book						
Mean	12.1	12.4	12.5	12.5	12.3	12.5
Median	12.0	12.3	12.0	12.0	12.5	12.5
Minimum	6.0	8.0	10.0	10.0	5.0	5.0
Maximum	18.0	18.0	15.0	15.0	15.0	17.5
No. in Sample	22	22	9	9	31	31
Book of Readings						
Mean	12.3	12.8	10.0	10.0	12.5	12.5
Median	15.0	15.0	10.0	10.0	12.5	12.5
Minimum	3.5	7.0	10.0	10.0	10.0	10.0
Maximum	15.0	15.0	10.0	10.0	15.0	15.0
No. in Sample	6	6	2	2	5	5

Notes: Royalty data obtained from a questionnaire completed by 119 faculty members chosen by random means of selection in over 40 colleges and universities and covering 178 hardcover books, many of which served multiple purposes, for example, an undergraduate text and a specialized book for sale to the trade. Rates are based upon books published since 1970. The Number in Sample refers to the number of books in the category, including those serving other purposes as well.

deed, at levels of 5,000 and 10,000 copies, relatively few texts carried a rate as low as 10 percent. Despite the significant economic difference between list price and either gross or net proceeds, 15 percent was both the median and modal royalty rate for textbooks. The difference between the lowest and highest rate reported in our survey generally ranged between 8 and 10 percentage points for any given royalty base.

When compared with textbooks, technical or specialized books had proportionately more low royalty rates and proportionately fewer high rates. Table 15 reveals striking differences. For example, if we look at the higher end of the royalty rate spectrum, about

30 percent of the specialized or technical books had royalty rates of 15 percent or higher, while this was true of 85 percent of the undergraduate texts.

Books with royalty rates higher than 15 percent at the 5,000th copy were uncommon. At this level of sales, only 1 book in 10 exceeded the rate of 15 percent, and almost all of these higher-rate books were classified as texts. For example, of the 62 books categorized as "technical or specialized book for a professional audience," only one carried a royalty rate above 15 percent. The limited data suggest that with sales of 5,000 copies, authors in economics and management are not likely to get a royalty rate exceeding 15

TABLE 15—CUMULATIVE PERCENTAGE DISTRIBUTION OF ROYALTY RATES FOR THE 5,000TH COPY BY TYPE OF BOOK AND ROYALTY BASE

Royalty Rate	Undergraduate Text	Graduate Text	Specialized or Technical
List Price			
17.0% and above	6.5%	4.3%	4.5%
15.0% and above	87.1	78.3	31.8
13.0% and above	87.1	78.3	36.4
11.0% and above	90.3	78.3	54.5
9.0% and above	100.0	95.7	95.5
3.0% and above	100.0	100.0	100.0
Books in Sample	31	23	22
Net Proceeds			
17.0% and above	27.5%	20.0%	0.0%
15.0% and above	82.5	73.3	32.3
13.0% and above	82.5	73.3	32.3
11.0% and above	92.5	93.3	71.0
9.0% and above	100.0	100.0	93.6
3.0% and above	100.0	100.0	100.0
Books in Sample	40	30	31

percent unless the book is a text. Even in the case of texts, only 15 percent of the books achieved these higher rate levels.

It is, of course, to be expected that those royalty rates which are based on net proceeds will be higher than those calculated on list price, since net proceeds are always smaller than list price. The data in Table 14 indicate that this sort of pattern does in fact occur, with some exceptions (for example, specialized or technical books). But the difference between the two mean rates is very small and is probably insufficient to compensate for the difference in magnitude of the two bases. In other words, the author seems to fare better when royalties are calculated on list price than when payments are based on net proceeds.

This impression was reinforced by the figures in Table 15 and others derived from the disaggregated data. We compared the distribution of texts by royalty rates for list price with the distribution calculated on equivalent rates for net proceeds (Table 16). The reader may recall that the discount to bookstores is commonly 20 percent for educational texts, and this figure gave us a first approximation for net proceeds (i.e., net proceeds equal to 80 percent of list). For example, with net proceeds averaging 80 percent

of list, a 15 percent royalty rate calculated on list price converts to 18.75 percent on net proceeds.

If there is no selection bias—that is, one type of base does not attract a disproportionate share of books perceived as highly saleable texts—and if the market makes a close adjustment for differences in the rate base, the percentage of undergraduate texts with an 18.75 percent rate on net proceeds should approximate the proportion with 15 percent on list price. But the proportion does not come close. For example, the 13 percent of the undergraduate texts paying the authors the converted rate of 18.75 percent on net proceeds compares very unfavorably with the 87 percent paying 15 percent on list price. Since frequencies seem to cluster at selected rates, we thought that use of a nearby cluster-point might reduce the gap appreciably. However, that turns out not to be so. A quarter of the books with net proceeds as the base achieved a royalty rate of 18 percent or higher, but this is far from the 87 percent with 15 percent on list price.

Farther down the royalty rate scale at 12 percent on list price (equivalent rate of 15 percent on net proceeds), market forces—or publisher's practices—seem to have eliminated most of the differential. The data sug-

TABLE 16—COMPARISON OF THE DISTRIBUTION OF UNDERGRADUATE AND GRADUATE TEXTS BY ROYALTY RATES WHEN ROYALTIES ARE CALCULATED ON LIST PRICE AND ON NET PROCEEDS<sup>a</sup>

Rate on List Price	List Price	Net Proceeds (80% of List Price)			
	Percent of Books with Rate or Higher	Equivalent Rate on Net Proceeds	Percent of Books with Equivalent Rate or Higher	Close Approx. Equivalent Rate	Percent of Books with Approx. or Higher
Undergraduate Texts					
18.00%	3.2%	22.50%	—	22.00%	—
15.00	37.1	18.75	12.5%	18.00	25.0%
12.00	90.3	15.00	82.5	15.00	82.5
10.00	100.0	12.50	87.5	12.00	92.5
8.00	100.0	10.00	100.0	10.00	100.0
Total Books	40		31		31
Graduate Texts					
18.00	4.3	22.50	—	22.00	—
15.00	78.3	18.75	6.7	18.00	16.7
12.00	78.3	15.00	73.3	15.00	73.3
10.00	95.7	12.50	90.0	12.00	93.3
8.00	95.7	10.00	100.0	10.00	100.0
Total Books	23		30		30

<sup>a</sup>Royalty rates on the 5,000th copy.

gest that for some books with net proceeds used in the calculation of royalties, the rate may be raised about enough to offset the effect of the smaller base, but for many others at the upper end of the spectrum (with rates of 15 percent or higher on net proceeds) the adjustment may be meager or nil.

The disaggregated data indicate that there is clustering of frequencies at selected rates, a practice which may be worth a comment. Of the 178 separate books for which there are complete data, only 4 had a royalty rate between 15 and 18 percent. Rates clustered at 15 percent and at 18 and 18.75 percent, with virtually no cases in between. This distribution suggests that publishers concentrate royalty rates at selected points. If this is indeed the case, the practice may suggest that authors who are offered 15 percent will find it difficult to inch their rate up to 16 or 17 percent. The practice may suggest further that authors trying to negotiate a rate higher than the 15 percent originally offered (particularly those with net proceeds as the base) may want to consider jumping to a counteroffer of 18.75 or 18 percent. From the standpoint of author's strategy, once a publisher

rejects an author's proposal of 16 or 17 percent, it may prove difficult to suggest successfully a still higher rate, whereas if one wants to make a second try, one can drop a point or two on the rate ladder.

Until now, our report has dealt with the basic royalty rate structure. In addition, most contracts call for different terms on various special types of sales. The categories most commonly covered include foreign sales, special editions, and disposal of overstock. The contract may also include clauses relating to other special circumstances, such as sales through radio or coupon advertising (probably not very relevant to this group of authors) and reduced royalties when rates of sale drop below a specified point.

#### Other Contractual Provisions

Beyond the contract terms dealing directly with author's earnings related to book sales, there are several other considerations which must be noted.

*Advances.* Many contracts provide that the publisher will pay to the author a fixed amount of money at one or several specified

dates, before royalties have been received and often before the manuscript has even been delivered. These payments are subsequently returned to the publisher, who retains future royalties until (if ever) all advances have been repaid. Such advances on royalty payments are an important benefit to the author. They shift part of the risk of the publication from the author to the publisher—if royalty payments never add up to the amount of the advance, the difference is simply borne by the publisher. Moreover, the author, by receiving his money earlier, if he wishes, can bank it sooner and earn additional interest. It is noteworthy that if a publisher decides, on grounds of quality, not to publish a book for which he has signed a contract before he had seen the complete manuscript, he often does not ask for a return of his advance payments.

In the previous survey, it was found that 27 percent of the contracts for hardcover books involved advance payments ranging from \$50 to \$6,000, with the median advance payment (for books on which any advance was offered) amounting to \$580.

The information on advances provided by the new sample is displayed in Table 17. In contrasting this new information, one should keep in mind the difference in samples and the effect of inflation.

The questionnaire asked for information both on advances offered and received. There was a slight difference (within a percent or two). Table 17 indicates advances actually received.

*Payments for Subsidiary Rights.* Of concern to the authors of technical research and textbooks are the arrangements concerning foreign sales and translations and rights to quote. Foreign sales revenues are usually divided between the author and publisher.

The author should certainly beware of any blanket provision assigning subsidiary rights to the publisher which includes no general formula for the sharing of proceeds. At the very least, the author may want the protection of a general clause, for example, the author will be paid 50 percent of the net amount actually received by the publisher, except as otherwise provided.

TABLE 17—ADVANCES

	Economics Departments	Business Schools
None	76%	73%
\$1–\$999	4	3
\$1,000–\$1,999	4	8
\$2,000–\$2,999	4	4
\$3,000–\$3,999	4	3
\$4,000–\$4,999	1	2
\$5,000 and over	7	7

### *Right to Quote, Abridge, and Reproduce.*

Where the publisher controls the right to permit others to quote from the work (frequently done on the ground that it helps promote sales) and to extract, digest, condense, abridge, and reprint, the publisher is under no contractual obligation to consult with the author on questions relating to this use of the work. While some publishers will routinely refer these questions to the author, others do so only intermittently or not at all. It should be pointed out that consultation with the author is a chore; sometimes the author cannot readily be located and, in any case, consultation imposes additional correspondence on the staff of the publishing firm and thereby increases its cost of doing business. In the event that the author is uncooperative in granting permission, consultation may also serve to reduce the book's visibility which, in turn, affects sales and possibly also the net proceeds that could be derived from the licensing of subsidiary rights. In any event, excessively rigid consultation requirements may constitute an undesirable obstacle to the diffusion of knowledge. On the other hand, the author may find that, if his writings are reprinted in whole or in part without his permission, they are taken out of context and used to promote causes which he opposes, or to defend positions with which he does not concur. Quotation of the work of a writer may tend to associate him indirectly, if not overtly, with the views held by the person who quotes him. Still another problem may arise where, in the process of abridgement or condensation, passages of critical importance in the author's eyes are

omitted or emasculated. The scholar also may want his writings brought up to date before they are reprinted. Unless the work is reissued principally as a matter of historical interest, he may feel a moral obligation to call attention, through footnotes at least, to important new developments or new research which may affect his previous conclusions. In all of these matters he has a professional reputation to protect.

Specific contract provisions requiring the publisher to consult with and receive the approval of the author (in effect granting the author veto power) might impose an unreasonable burden on the publisher. But an alternative and less rigid approach to the problem has much to be said for it. The author may well find it useful to protect his interests by asking for the insertion in his contract of a somewhat flexible clause on the subject. Such a clause would require that when the author's work is to be quoted at length, abstracted, condensed, digested, abridged, or reprinted the publisher will make reasonable efforts to consult the author before granting permission.

*Number of Free Copies and Price of Additional Copies.* Most contracts provide that the author will receive a number of copies of his book when it appears. The Baumol-Heim study reported that the number of free copies usually ranges from 6 to 10, with these two extreme figures occurring most frequently. Table 18 indicates a somewhat different response.

The author is usually also entitled to purchase at a discount as many additional copies of the book as he desires for his own use. The discount on these purchases varies quite widely, usually ranging from 20 to 40 percent of the list price. Though it is not specified explicitly, presumably the author also receives royalty payments on the copies he purchases at a discount, since the discount would appear to consist largely of retailer's margin which is avoided by the elimination of the middleman.

*Pricing.* In their study, Baumol and Heim observed that the contract normally leaves the pricing decision entirely in the hands of the publisher, who need neither

TABLE 18—FREE COPIES

Number of Free Copies	Economics Departments	Business Schools
2	3	1
3	1	1
5	10	4
6	25	15
7	—	1
8	5	3
10	58	37
More than 10	29	29
No response	46	30

explain nor even inform the author of his decision. Since most contracts are signed well before the final manuscript is delivered, it is of course not possible for the publisher to commit himself to a selling price for the book at that time. However, it might be desirable for the publisher to explain the nature of his pricing calculations to the author, particularly since this might induce the author to take more careful account of pricing costs as he works. In any event, the author should be informed of the pricing decision as soon as it is made. Many publishers provide to the author soon after publication information on price, and number of copies printed, promotion plans, review copies sent, etc. This can help cement relations between author and publisher, and frequently the author is able to make useful suggestions on the basis of this information. And where the author is induced to accept unusually low royalties or to provide a publishing subvention as a means to keep down the retail price of the volume, the publisher should be required to indicate in writing exactly how the retail price will be affected. In this way much unpleasantness and some abuses can be prevented.

Inflation has, of course, had a considerable effect on book prices. Yet, the author of even a research volume who hopes for its use as a graduate text must be concerned when, as occurred recently, a North-Holland book on game theory was priced at over \$100.

*Cost of Corrections.* Contracts usually include a clause protecting the publisher from

the cost of excessive corrections in proof. On rare occasions, the author is required to bear the cost of any corrections other than those occurring as a result of the printer's errors. In some cases, the author is asked to pay half the cost of corrections, with the publisher's liability limited to some fixed amount, ordinarily \$100 or \$200. More commonly, the author is permitted to supply corrections in proof at no cost to himself so long as they do not exceed 5 to 10 percent of the entire cost of setting the volume into type. In a few contracts, this percentage may be raised to 15 or even 20 percent. Since corrections may cost as much as \$1.50 a line, they can well constitute a major expense for the publisher, and, therefore, for the author.

*Preparation of the Index.* It appears that for the most part the author either prepares the index or it is paid for out of royalties. However in the bargaining process, the publisher will sometimes agree to bear this cost.

*Options on Subsequent Writings.* A contract often commits the author to submit his new book (or sometimes his next two books) to the same publisher, who has the option of undertaking publication on terms that are "mutually agreeable" after "negotiation in good faith." It is not clear what choices are open to the author if some other publisher offers him rates on a subsequent manuscript higher than those proposed to him under the option contract.

Most option clauses require the publisher to indicate that he will exercise his option, that is, that he will let the author know whether he is willing to publish the book within 60 or sometimes 90 days after the manuscript has been received. If the author has not heard from the publisher by that time, he is free to submit his manuscript elsewhere. The absence of such a time limit in the option clause can sometimes constitute a real disadvantage to the author because it can result in a very substantial delay in publication of his book. The author is therefore well advised to check for the presence of such a time limit in any option provision.

In this survey, only one individual reported on a publisher's first refusal clause for his next book.

*Information on Reprinting.* It is often desirable to have the opportunity to correct errors before reprinting. This requires that the author be aware of the reprinting. Table 19 indicates that communication between the

TABLE 19—INFORMATION ON REPRINTING

	Economics Departments	Business Schools
Informed	96	72
Not informed	40	38
No response	46	17

TABLE 20

	Economics Departments		Business Schools	
	Short MS Range	Long MS Range	Short MS Range	Long MS Range
Detailed Reading				
Maximum	\$100-500	\$200-500	\$150-300	\$350-800
Average	75-300	75-300	50-400	300-500
Minimum	50-200	80-350	75-150	100-300
Close Reading				
Maximum	\$50-500	\$100-400	\$150-300	\$300-700
Average	120-400	120-300	100-250	—
Minimum	80-300	80-100	150-200	—
Cursory Reading				
Maximum	\$50-300	\$50-200	\$75-250	\$100-450
Average	50-200	80-150	40-175	75-125
Minimum	50-100	50-60	25-150	50-100

TABLE 21—ACADEMIC RANK

	Professor	Associate Professor	Assistant Professor	Lecturer
Economics Departments	78	17	1	—
Business Schools	50	14	4	1

TABLE 22—NUMBER OF BOOKS PUBLISHED

Authors in:	1	2	3	4	5	6	7	8	9	10 or more
Economics Departments	24	20	11	4	10	8	3	0	3	9
Business Schools	21	15	12	7	5	4	0	0	1	3

authors and publishers on this subject was quite poor.

*Manuscript Reviewing.* Especially for research and technical books peer review of manuscripts is of importance. Table 20 indicates the ranges in payment for a detailed, close, and cursory reading of a nontechnical but specialized manuscript. (There were 44 responses from the business schools and 46 from economics departments.)

There appear to be no particular surprises, or, for that matter, exploitation in prices paid for manuscript reading. One should also note that to many of the readers there is an external benefit to be had inasmuch as some of the reading is of professional interest. Yet reading fees probably work out at no more than \$200–\$300 a day at best, which is far from consulting fees.

#### Profile of Authors' Rank and Number of Publications

As can be seen from Table 21, both at the economics departments and business schools, book publication is reported primarily by full professors. The number of books pub-

lished by the authors is indicated by Table 22, showing a mode at 1 book per author, but with a tail going beyond 10 or more.

#### On Author-Publisher Relations in Practice

Our survey was, in essence, consistent with the Baumol-Heim survey, and the interested reader is referred to pages 39–43 of their article. On the whole, the authors appear to be reasonably satisfied and reasonably well served. The major problem appears to be confusion concerning royalty payments, lack of knowledge of the authors about publishing, and poor communication between authors and the press.

According to Paul Kingston, the median writing-related income of American authors in 1979 was \$4,775. Those who worked 40 or more hours a week at genre fiction had a median income from writing of \$31,500. For those economists whose prime reason for working is money, teaching EC 101 appears to be a better-paid proposition than most writing, except for the production of the basic textbook for EC 101.

AMERICAN ECONOMIC ASSOCIATION

---

PROCEEDINGS  
OF THE  
NINETY-FIFTH  
ANNUAL  
MEETING

NEW YORK, NEW YORK  
DECEMBER 28–30, 1982

## Minutes of the Annual Meeting New York, New York December 29, 1982

The Ninety-Fifth Annual Meeting of the American Economic Association was called to order by President Gardner Ackley at 9:37 P.M., December 29, 1982, in the East Ballroom of the New York Hilton Hotel. The minutes of the meeting of December 29, 1981 were approved as published in the *American Economic Review, Papers and Proceedings*, May 1982, pages 399-400.

The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), Managing Editor of the *Journal of Economic Literature* (Moses Abramovitz), and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports which were distributed at the meeting. (See their reports, and that of the Managing Editor of the *American Economic Review*, published elsewhere in this issue.) The Treasurer reported that the Executive Committee's decision to continue AEA membership in the Consortium of Social Science Associations would require an additional \$35,000 in expenses which was not reflected in the 1983 budget contained in his written report.

Three members raised questions concerning the finances of the Association. David Fritz noted that *Job Openings for Economists* (*JOE*) has continually run a deficit and asked why. The Treasurer and the Secretary responded that the Association's auditors had judged *JOE* to be an "unrelated business activity" and, consequently, subject to the income tax. Much of the expense of *JOE* was allocated overhead rather than direct costs. In an economic sense, there was probably not a deficit. William Vickrey questioned the wisdom of the Association's willingness to charge high subscription fees to libraries, which already have significant financial problems. The Treasurer responded that the subscription rate was not high compared to other journals' rates, a higher rate might be justified on the basis of multiple usage of library copies, and the Executive Committee might be willing to reduce the rate if other

journals reduced theirs, but to do so unilaterally would not help the libraries much and would have a significant impact on the Association's finances. In any case, the rate structure was being reviewed because of a recent adverse ruling by a post office auditor who found that the rates did not meet postal requirements for second class mailing privileges. Donald Dalton inquired as to how the Association could earn such significant "Investment Gains" since 1981. The Treasurer pointed out that real capital gains (or losses) are recognized whether realized or not, and that each year's gain from equities is spread over a three-year period. The projected 1983 figure includes one-third of the gain on equities during 1981, one-third of the gain in 1982, and one-third of the projected 1983 increase. The audited financial statements, which appear in the June issue of the *AER*, contain footnotes that explain the procedure used in recognizing income from the portfolio.

In response to a question about the status of the 1978 *Index*, Naomi Perlman, Associate Editor of the *JEL*, responded that she expected it to appear in the spring of 1983. Labor shortages and work-load difficulties had caused the delay. Abramovitz elaborated by calling attention to the paragraphs in his written report related to the quantity of documentation services provided by the Pittsburgh offices of the *JEL*, including an arrangement with Dialog Information Retrieval Service to provide online computer access (in early 1983) to the *JEL* and *Index* data bases.

The Secretary then presented the following resolution, which was adopted unanimously and with applause:

BE IT RESOLVED that this meeting record a special vote of appreciation for the members of the 1982 Allied Social Science Associations' Convention Committee, chaired by Peter

Fousek, and an extra special "thank you" to Alice Christensen who has cochaired her second ASSA meeting with efficiency and style.

Alfred Kraessel and Walter D'Ull submitted the following resolution to the Secretary thirty days prior the meeting:

Re: Examination of economic illiteracy problem, it is moved that the AEA appoint or elect a committee for the purpose of examining the problem of investigating economic illiteracy and misinformation and also consider participation in accreditation procedures of institutions of higher learning by professional and regional accreditation associations for the purpose of maintaining adequate standards in the instruction of economics.

Kraessel was recognized by the President and spoke in favor of the motion. He deplored the state of economic literacy in the country and attributed it in part to poor

working conditions, large classes, low status of economists, and a decline in classical education in general. He advocated the Association's involvement in accreditation as a potential solution. Fels pointed out that the Association already has a Committee on Economic Education that is concerned with economic illiteracy. Charles Schultze spoke against the part of the resolution that would move the Association toward an accrediting procedure. Andrew Brimmer offered a substitute motion to refer the literacy problem to the Committee on Economic Education and eliminate the charge concerning accreditation. It was voted to accept the substitute motion. The main motion, now the substitute motion, was then approved.

Ackley then introduced W. Arthur Lewis, the President of the Association for 1983, to the assemblage.

The meeting was adjourned at 10:15 P.M.

Respectfully submitted,

C. ELTON HINSHAW, *Secretary*

## Minutes of the Executive Committee Meetings

### Minutes of the Meeting of the Executive Committee in New York, New York, March 26, 1982.

The first meeting of the 1982 Executive Committee was called to order at 9:10 A.M. on March 26, 1982 in the Sutton Room of the New York Hilton Hotel, New York, New York. Members present were Gardner Ackley (presiding), Moses Abramovitz, Elizabeth Bailey, William Baumol, Robert Clower, Robert Dorfman, Rendigs Fels, Ann Friedlaender, C. Elton Hinshaw, Alfred Kahn, W. Arthur Lewis, Robert Lucas, and Joseph Stiglitz. Leo Raskind, counsel of the Association, was also present. Members of the Nominating Committee present for part of the meeting were Marcus Alexis, Charles C. Holt, Daniel J. B. Mitchell, Steven Salop, Isabel V. Sawhill, and Robert M. Solow. Present as a guest for part of the meeting was Lloyd G. Reynolds.

*Minutes.* The second sentence of the paragraph labeled *1982 Program* in the minutes of the December 27, 1981 meeting was amended to read: Lewis has already received 60 to 70 proposals for contributed papers and expects the number to increase. With this amendment, the minutes were approved.

*Report of the Secretary* (Hinshaw). The Secretary reported the current schedule for future annual meeting sites: New York (1982), San Francisco (1983), Dallas (1984), and New York (1985). Registrations for the 1981 Washington, D.C. meetings totaled 6,340, the highest number since the 1977 meeting in New York. He also reported that both the German Economic Association and the Italian Society had accepted the arrangements offered to them concerning the purchase in bulk of issues of the *Journal of Economic Literature*. He raised the question of holding future Executive Committee meetings away from the East Coast. It was agreed to continue to meet in Washington or New York. It was also agreed that the first meeting of the Committee in 1983 would be on March 18th.

*Report of the Treasurer* (Fels). The Treasurer reported that, at the end of 1981, the

net worth of the Association was \$1,390 thousand; the 1981 surplus was \$331 thousand compared to \$84 thousand in 1980. The increase in the surplus resulted primarily from the increase in subscription prices from \$43 to \$100 effective January 1, 1981. The 1982 budget projects a surplus of \$131 thousand. He advocated continuing the policy of raising nominal dues less than the rate of inflation so that real dues will continue to fall. Between 1976 and 1982, dues in nominal terms rose 24 percent whereas the implicit consumption deflator went up 55 percent. Dues in real terms fell about 20 percent. It was VOTED to increase the base dues rate from \$31 to \$32 (3 percent). It was VOTED to allocate \$15,000 to the Committee on the Status of Women in the Economics Profession for 1982. It was VOTED to reimburse Marcus Alexis, chair of the Committee on the Status of Minorities in the Economics Profession, for his travel expenses to the 1981 annual meeting in Washington, D.C.

Concern has been expressed about the impact of the significant increase in subscription rates on the budgets of libraries. It was VOTED to discuss the issue of subscription rates with other Associations.

*Report of the Editor of the American Economic Review* (Clower). Clower reported that the December 1982 issue was complete and ready to publish; he was currently selecting articles for the March 1983 issue; the lag between acceptance of an article and its publication was now slightly over one year; and the number of manuscripts submitted had increased to about 800 or 900 per year. He requested permission to appoint an additional person to the editorial staff to help with the work and to serve as a second in command. There followed a wide-ranging discussion of the increasing burdens being placed on both the *AER* and *JEL* editors, and ways to help make the jobs more manageable, including moving to co-editors. It was VOTED to approve Clower's request (without precedent-setting implications) and to establish an *ad hoc* committee to review

the structure of the two editorial offices and consider alternatives, such as co-editors.

During the discussion, it was suggested that the Executive Committee might wish to adopt editorial guidelines for the journals. The following set of guidelines was circulated as an example:

1. The editorial policies of the Association's journals are defined by the statement of the objects of the society set forth in its Certificate of Incorporation.

2. In particular, the journals should encourage "perfect freedom of economic discussion." To this end it should be presumed that all papers submitted for publication are entitled to be reviewed for scholarly merit by competent, unprejudiced members of the profession, and all those judged meritorious should be eligible for publication.

3. In the event that papers eligible for publication are received at a greater rate than can be published, the editors should select the papers to be published by assigning priorities based on their interpretation of the strength of the reviewers' recommendations and their own judgment of the significance of the papers. Papers without sufficient priority to be published in any of the succeeding three issues should be returned to the authors with an explanation of the reason.

4. In spite of the presumption mentioned above, the editors may decline a paper without having it reviewed for any of the following reasons: (a) The paper does not deal with a topic that falls within the objects of the society; (b) The substance of the paper is known to have been published previously, either by the same author or by others; or, (c) The scholarly quality of the paper is clearly so inadequate that it is highly unlikely that unprejudiced, competent reviewers would recommend its publication.

It was VOTED to extend the charge of the committee appointed to review the structure of the editorial functions to include the issues of guidelines and procedures.

It was VOTED to offer a second three-year term as editor of the *AER* to Clower. The second term would end on December 31, 1986.

*Report of the Editor of the Journal of Economic Literature* (Abramovitz). Abramovitz confirmed that the duties and responsibilities of editing the *JEL* and managing the office were more onerous than he anticipated. He too needs help. It was VOTED to offer Abramovitz a second three-year term with the understanding that he would present a proposal for restructuring the managing editor's function. His second term would extend to December 31, 1986.

Raskind, counsel of the Association, was asked to prepare an amendment to the bylaws that would allow the Association to appoint more than one managing editor for each journal in case this becomes desirable.

*Report on the 1982 Program* (Lewis). The President-elect stated that the program was complete and called attention to three issues that had arisen during its planning: (1) The number of contributed papers has increased significantly. (2) Other member associations of Allied Social Science Associations are requesting that more of their sessions be jointly sponsored by the AEA. (3) One consequence of inviting "public" figures to chair sessions, give papers, or be commentators may be to deplete attendance at the more "scientific," scholarly sessions. In the discussion that followed, a variety of opinions was expressed concerning the merits (or lack thereof) of contributed paper sessions, joint sponsorship by the AEA of other associations' sessions, and "big name" sessions. Nothing was concluded.

*Soviet Exchanges* (Reynolds). Reynolds reviewed the history of the exchange program and, on behalf of the Committee on Soviet Exchanges, recommended that the program be started up again with a symposium to be held in the United States in 1983. It was decided to check with the National Academy of Sciences to see if its decision to suspend exchanges with the Soviet Union had been changed. Ackley was empowered to reinstate the exchange if the National Academy had begun to sponsor such programs again. Sentiment was expressed in support of revising the program to include East Europe and establish a Committee on Soviet and East European Exchanges.

*Nominating Committee* (Solow). Solow reported the following nominees for offices in the 1982 election: Vice Presidents (two to be chosen), Oliver Williamson, Juanita Kreps, Edmund Phelps, and Lloyd Ulman; Executive Committee members (two to be chosen), William Nordhaus, A. Michael Spence, Glenn Loury, and Thomas Sargent. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together, chose as nominee for President-elect, Charles Schultze, and as Distinguished Fellows, Joe Bain and Gerard Debreu.

*Ad Hoc Committee on J. B. Clark Medal* (Solow). Solow (reporting for Daniel McFadden, chair of the committee; Martin Feldstein was also a member) reviewed the Committee's recommendations. In addition to a suggested schedule for the selection process, the Committee recommended an election procedure. The procedure calls for the Honors and Awards Committee to bring a single nominee for the award to the Electoral College (the Honors and Awards Committee and the Executive Committee meeting together) with a ranked, short list of other candidates. Any member of the Electoral College can move to substitute another candidate for the person originally nominated (for example,  $x$  for  $y$ ). Additional substitute motions can be made (for example,  $z$  for the winner between  $x$  and  $y$ ). Substitute motions require a simple majority on a written ballot. The final nominee is approved if passed by a simple majority on a written ballot. If the final nominee does not receive a majority vote, the Clark medal will not be awarded. It was VOTED to adopt the recommended procedure.

There being no further business, the meeting adjourned.

**Minutes of the Meeting of the Executive Committee in New York, New York, December 27, 1982.**

The second meeting of the 1982 Executive Committee was called to order at 10:00 A.M. on December 27, 1982 in the Sutton Parlor of the New York Hilton Hotel, New York, New York. Members present were Gardner

Ackley (presiding), Moses Abramovitz, Elizabeth E. Bailey, William J. Baumol, Robert W. Clower, Robert Dorfman, Rendigs Fels, Ann F. Friedlaender, Robert J. Gordon, C. Elton Hinshaw, Alfred E. Kahn, W. Arthur Lewis, Robert Lucas, and Joseph E. Stiglitz. Leo J. Raskind, Counsel of the Association, was also present. Newly elected members of the 1983 Executive Committee present were Juanita M. Kreps, William D. Nordhaus, Edmund S. Phelps, Charles L. Schultze and A. Michael Spence. Present as guests for parts of the meeting were Marcus Alexis, Donald J. Brown, Naomi Perlman, Albert Rees, and Lloyd G. Reynolds. Ackley opened the meeting by expressing his and the Association's appreciation to the members whose terms were expiring for their services and welcoming the new members of the 1983 Executive Committee.

*Minutes.* The minutes of the March 26, 1982 meeting were approved as written and circulated.

*Ad Hoc Committee on Publications* (Rees). The Committee considered first the growing work load of the managing editors, next the rights of authors not to have materials arbitrarily rejected, and finally the need for a new journal to include types of materials not now in the *AER* or *JEL*, especially short articles or notes on current research and current issues. The Committee's suggestions for the Executive Committee's consideration along with their disposition are discussed below: (1) Substantially reduce in size the Board of Editors of the *AER* and pay members of this smaller board a modest honorarium. It was VOTED to allow the managing editor of the *AER* to reduce the size of the board to not less than six members; it was decided not to pay honorarium. (2) The Bylaws of the Association should not be amended to permit the selection of joint managing editors of the *JEL*; one person should be responsible to the Executive Committee for each journal. The Executive Committee did not specifically discuss this issue but apparently accepted the recommendation for no action was taken. (3) Divide the *JEL* into two journals, each with its own editor. One, the new *JEL*, would contain the

present articles and book reviews, plus possible additional material which might have appeared in a new, third journal. The second journal would contain the indexes, annotations, and abstracts of new books and current periodicals now in the "back part" of the *JEL*. After an extended discussion of the merits of a split *JEL* versus a unified journal, the organization and communication problems arising from coordinating the three offices (Stanford, Pittsburgh, and Nashville) involved in the production and distribution of the *JEL*, and related issues, the Executive Committee VOTED to appoint a new committee to examine the long-term arrangements needed for the journal, including its organizational structure and format, and search for a replacement for the managing editor who indicated he did not wish to complete his second term. The President was delegated the power to make any interim arrangements deemed necessary until the committee reports. (4) The "Guidelines for Editors of the Association's Journals" prepared for the Executive Committee meeting of March 27, 1982 (see the minutes of that meeting) should not be adopted nor should attempts to formulate similar ones be made; the basic protection of authors is the existence of many (at least 260) independently edited journals. The Executive Committee concurred. (5) Do not begin a new journal devoted to materials not now included in the *AER* or *JEL*. The Executive Committee showed no inclination to begin a new journal.

*Report of the Secretary* (Hinshaw). The Secretary reported that next year's annual meetings will be held in San Francisco on December 28-30, and the Placement Service will open on the 27th. The schedule for subsequent meetings is Dallas in 1984 and New York in 1985. He recommended New Orleans as the site for 1986. The 1978 poll of members ranked New Orleans third among the fourteen cities listed on the ballot which asked members to rate cities from one (excellent) to four (unsatisfactory) as sites for future meetings (San Francisco was first, Boston second). The results of the most recent poll of members is shown in Table 1. New Orleans is clearly the preferred city among

TABLE 1—INFORMATIONAL BALLOT CONCERNING THE ANNUAL MEETING OF THE AEA

I. Time of Year—Vote for One	Votes	Rank
Between Christmas and New Year	2,363	2
Weekend Immediately after New Years	3,563	1
Total	5,899	
II. Site Selection—Vote for the two cities you most prefer in each area		
Area I:		
Washington, D.C.	3,080	1
Boston	2,628	2
New York	2,497	3
Montreal	2,067	4
Philadelphia	1,191	5
Atlantic City	752	6
Area II:		
New Orleans	3,344	1
Chicago	2,338	2
Toronto	1,737	3
Miami Beach	1,717	4
Atlanta	1,650	5
St. Louis	824	6
Area III:		
San Francisco	4,550	1
San Diego	2,621	2
Los Angeles	1,682	3
Dallas	1,240	4
Houston	971	5
Kansas City	934	6
Total Ballots	6,397	

those listed in Area II and second to San Francisco in total votes received. We last met in New Orleans in 1971. It was VOTED to accept the Secretary's recommendation.

On both the 1978 and 1982 ballots, members were asked about preferred dates for the meetings. In 1978 members were asked to rank six options. "Between Christmas and New Year's" ranked one, "Immediately after New Year's" second. On the recent ballot, members were asked to vote for one or the other of these two dates. "After New Year's" garnered 60 percent of the votes.

The Regency Inn sued the Allied Social Science Associations for breach of contract after the Denver meetings in 1980. The case was tried this summer, and the Judge found in the Regency's favor. The decision is being appealed.

The Executive Committee at its meeting on March 8, 1974, voted to require that, to be considered at the annual business meeting, proposed resolutions must be submitted

to the Secretary at least one month in advance. At the December 29, 1981 annual business meeting, Alfred Kraessel urged the Executive Committee to review the policy; he found business meetings under the old policy more exciting. The Secretary noted that he prefers short, dull business meetings. It was decided to retain the rule.

JAI Press Inc. has proposed to prepare with AEA cooperation and approval, an author, title, and subject cumulative index for the *American Economic Review*, covering volumes 1-72. It was VOTED to allow JAI Press to prepare such an index, but not to act as a sponsor. Appropriate royalties should be charged if AEA copywriters are involved.

*Report of the Treasurer (Fels).* The Treasurer reported that a substantial surplus is in prospect for this year and a smaller surplus of about \$160,000 in 1983 in spite of an estimated 10 percent increase in expenditures. (See elsewhere in this issue for his written report.) With inflation now running at an annual rate of about 5 percent, the 3 percent increase in dues that goes into effect on January 1, 1983, means a small dues reduction in real terms, a continuation of the policy begun in 1976. Since January 1, 1976, real dues have declined on the order of 20 percent. He also reported that we may have to restructure the dues and subscription rates because of postal regulations; a postal authority has audited the Association's compliance with rules governing second class mailing privileges and has initially determined that the dues structure does not meet postal requirements. It was VOTED to approve the 1983 budget. It was also VOTED to approve the travel expenses to the 1982 annual meetings of one of the members of the Executive Committee. It was decided to review the current policy of not reimbursing Executive Committee members' travel expenses to the annual meetings.

*Report of the Managing Editor of the American Economic Review (Clower).* After reviewing his written report (see elsewhere in this issue), Clower informed the Executive Committee that he had no desire to continue as Editor beyond his second term which ends December 31, 1986. Given the amount of time necessary to find replacements, he

thought such early notification would be helpful.

*Report of the Managing Editor of the Journal of Economic Literature (Abramovitz).* The Editor reviewed his written report (see elsewhere in this issue). It was VOTED to approve his nominations of Robert Havenman and John Panzar to three-year terms on the Board of Editors.

*Report of the Director of Job Openings for Economists (Hinshaw).* See elsewhere in this issue for his written report.

*Committee on U.S.-Soviet Exchanges (Reynolds).* Reynolds reported that the Seventh U.S.-Soviet Economics Symposium, will be held in the United States in 1983 on the subject "The Economics of Non-Renewable Resources." The meeting is scheduled for New Haven, Connecticut, on June 4-6, 1983. After the symposium, the Soviet delegation will as usual visit New York, Washington, and probably one other city over a two-week period. He expects a delegation of ten Soviet economists. Fourteen American economists have been invited. The 1983 meeting will presumably be followed by an invitation for a delegation of U.S. economists to visit the U.S.S.R. for the Eighth Symposium in 1984, completing the fourth "round" of the exchange. The Committee was encouraged to explore a different type of East-West interchange. This would consist of one or more relatively small meetings involving economists from Eastern and Western Europe as well as from the United States and the U.S.S.R. Discussion would focus on topics of general interest to economists in industrialized countries, such as enterprise management, productivity, invention and innovation, and technical progress. Concern was expressed about the possible conflict between our desire to encourage academic freedom in these countries and our willingness to continue to arrange exchange programs.

*1983 Program (Schultze).* Schultze reported that the invited sessions had been arranged; that is, chairmen had been appointed who had agreed to organize sessions dealing with the specific topics suggested. The program will center around recent developments in micro theory and their implications for macroeconomics. Relatively few

contributed paper sessions had yet been organized.

*Committee on the Status of Minority Groups* (Alexis). Alexis reported that 1982 was the last year the summer program would be held at Yale. The Association owes a debt of gratitude to Yale and Donald Brown, the Director of the program while it was at Yale, for a magnificent job. The scope of the activities of the Committee has increased to include providing financial aid to students attending graduate school. Acting on the Committees' recommendation, the Executive Committee VOTED to approve the move of the program to the University of Wisconsin for 1983-85. It was also VOTED to allocate \$10,000 toward the program's support. Alexis reported also that he understood that the Sloan Foundation will give a \$300,000 grant to the existing program and possibly \$100,000 for a second, smaller program especially for Hispanics. The University of Southern California was interested in hosting such a program.

*Committee on the Status of Women in the Economics Profession* (Bailey). Bailey gave highlights from her written report (see elsewhere in this issue). It was VOTED to allocate \$15,000 in support of the Committee's work.

*The Consortium of Social Science Associations*. Henry Aaron, AEA representative to COSSA, wrote a letter advocating our continued participation. After a discussion of the activities of COSSA and an expressed desire for more detailed information on the

benefits of membership, the Executive Committee VOTED to contribute \$35,000 and continue our membership for 1983.

*Simmons College Proposal*. Representatives from Simmons College submitted a proposal to enlist the support of the AEA in a joint venture: to prepare a grant proposal entitled "Economics Career Resources Project," a program to be organized and administered by the Simmons College Libraries, with the Association's formal cooperation and sponsorship. The project would develop multimedia career information packages and provide for their distribution. It was VOTED to be helpful and encouraging, but not formally to sponsor the project.

*Other Business*. Amnesty International had written asking the Association to adopt a prisoner of conscience. It was decided to acquire more information before taking any action. Thomas F. Hady had written advocating graduating dues simply according to income. It was decided to take no action until the impact on the dues structure of the postal authorities findings is determined. The American Sociological Association sought the AEA's endorsement of a resolution concerning the revocation of advanced degrees held by those who apply to emigrate from the Soviet Union. No action was taken.

The meeting adjourned at 4:42 P.M.

Respectfully submitted,  
C. ELTON HINSHAW, *Secretary*

## Report of the Secretary for 1982

*Annual Meetings.* In 1983 the annual meetings will be held at the San Francisco Hilton Hotel in San Francisco on December 28–30. The schedule for subsequent meetings is December 28–30, 1984 in Dallas, December 28–30, 1985 in New York, and 1986 in New Orleans. Employment services will be provided at the annual meeting beginning December 27.

*National Registry.* The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision of the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and their professional obligation to list their openings.

*Membership.* The total number of members and subscribers, shown in Table 1, reached an all-time high of 26,787 at the end of 1975. After declining for two years, the total increased in 1978, 1979, and again in 1980. There was a small decline in 1981 and virtually no change in 1982.

*Permission to Reprint and Translate.* Official permissions to quote from, reprint, or translate and reprint articles for the *American Economic Review* and the *Journal of Economic Literature* totaled 230 in 1982 compared to 299 in 1981. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to get the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of \$150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

TABLE 1—MEMBERS AND SUBSCRIBERS  
(End of Year)

	1980	1981	1982
Class of Membership			
Annual	16,219	16,738	16,771
Junior	1,811	1,800	1,895
Life	383	385	384
Honorary	33	32	32
Family	331	374	397
Complimentary	624	607	607
Total Members	19,401	19,936	20,086
Subscribers	7,094	6,291	6,171
Total Members and Subscribers	26,495	26,227	26,257

*Elections.* In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of Robert M. Solow, Chair, Marcus Alexis, Charles C. Holt, Daniel J. B. Mitchell, William D. Nordhaus, Steven Salop, and Isabel V. Sawhill submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

*President - Elect*  
Charles L. Schultze

<i>Vice President</i> Juanita M. Kreps Edmund S. Phelps Lloyd Ulman Oliver E. Williams	<i>Executive Committee</i> Glenn C. Loury William D. Nordhaus Thomas J. Sargent A. Michael Spence
--	---

The Secretary prepared biographical sketches of the candidates and distributed

ballots last summer. On the basis of the canvass of the ballots, I certify that the following persons have been duly elected to the respective offices:

*President-Elect* (for a term of one year)

Charles L. Schultze

*Vice-President* (for a term of one year)

Juanita M. Kreps

Edmund S. Phelps

*Executive Committee* (for a term of three years)

William D. Nordhaus

A. Michael Spence

Number of legal ballots	6,245
Number of invalid envelopes	314
Number of envelopes received after Oct. 1	71
Number of envelopes returned	6,630

*AEA Staff.* Mary Winer, Administrative Director, Kimberly Adair, Norma Ayres, Ersye Burns, Ettamene Byrd, Marcia McGee, Violet Kohsman, Dale Wagner, and Jacquelyn Woods handle the day-to-day operations of the Association. I wish to thank them for their efficient and dedicated work.

*Committees and Representatives.* Listed below are those who served the Association during 1982 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they have been appointed. On behalf of the Association, I wish to thank them all for their services.

*Ad Hoc Committee on Relations with IEA*

Fritz Machlup, *Chair*

Anne O. Krueger

Franco Modigliani

Paul A. Samuelson

C. Elton Hinshaw, *ex officio*

*Ad Hoc Committee on Publications Policy*

Albert E. Rees, *Chair*

George H. Borts

Rendigs Fels

Ann F. Friedlaender

Robert J. Lampman

Joseph A. Pechman

*Ad Hoc Committee on John B. Clark Award*

Daniel McFadden, *Chair*

Robert M. Solow

Martin S. Feldstein

*Ad Hoc Committee on Publishing Contracts*

Martin Shubik, *Chair*

Peggy Heim

Leo J. Raskind

C. Elton Hinshaw, *ex officio*

*Budget Committee*

Rendigs Fels, *Chair*

Martin S. Feldstein (1982)

Elizabeth E. Bailey (1983)

Ann F. Friedlaender (1984)

Gardner Ackley, *ex officio*

W. Arthur Lewis, *ex officio*

*Census Advisory Committee*

Norman J. Simler, *Chair* (1983)

Carolyn Shaw Bell (1982)

Ronald L. Oxaca (1982)

Walter E. Williams (1982)

Ann D. Witte (1982)

Arnold Zellner (1982)

Michael K. Evans (1983)

Victor R. Fuchs (1983)

Zvi Griliches (1983)

Sherwin Rosen (1983)

Morris A. Adelman (1984)

Rosanne Cole (1984)

Martin H. David (1984)

Sidney L. Jones (1984)

Edwin Mansfield (1984)

*Committee on Economic Education*

Allen C. Kelley, *Chair* (1982)

Campbell R. McConnell (1982)

George L. Bach (1983)

Marianne A. Ferber (1983)

Herbert Stein (1983)

W. Lee Hansen (1984)

Michael K. Salemi (1984)

John J. Siegfried (1984)

Rendigs Fels, *ex officio*

*Economics Institute Policy and Advisory Board*

Edwin S. Mills, *Chair* (1986)  
 Axel Leijonhufvud (1982)  
 Dwight H. Perkins (1983)  
 G. Edward Schuh (1983)  
 Bent Hansen (1984)  
 Louis T. Wells (1984)  
 Robert E. Evenson (1985)  
 W. Lee Hansen (1985)  
 John R. Moroney (1986)

*Finance Committee*

Rendigs Fels, *Chair*  
 Robert G. Dederick (1982)  
 Robert J. Genetski (1983)  
 Sidney Davidson (1984)

*Committee on Honorary Members*

Richard A. Musgrave, *Chair* (1986)  
 Hollis B. Chenery (1982)  
 Tibor Scitovsky (1982)  
 Hendrik S. Houthakker (1984)  
 George J. Stigler (1984)  
 Hal Varian (1986)

*Committee on Honors and Awards*

Anne O. Krueger, *Chair* (1983)  
 Carl F. Christ (1982)  
 Dale T. Mortensen (1983)  
 Daniel McFadden (1985)  
 Oliver E. Williamson (1985)  
 Robert Eisner (1987)  
 William Vickrey (1987)

*Nominating Committee* (1982)

Robert M. Solow, *Chair*  
 Marcus Alexis  
 Charles C. Holt  
 Daniel J. B. Mitchell

William D. Nordhaus  
 Steven Salop  
 Isabel V. Sawhill

*Committee on Political Discrimination*

Martin Bronfenbrenner, *Chair* (1982)  
 Lester C. Thurow (1982)  
 Herbert Gintis (1983)  
 Richard R. Nelson (1983)  
 Harold C. Barnett (1984)  
 Anne P. Carter (1984)

*Committee on the Status of Minority Groups in the Economics Profession*

Marcus Alexis, *Chair* (1982)  
 Richard Freeman (1982)  
 Gerald D. Jaynes (1982)  
 Glenn C. Loury (1982)  
 Vincent McDonald (1982)  
 Jeffrey G. Williamson (1983)

*Committee on the Status of Women in the Economics Profession*

Elizabeth E. Bailey, *Chair* (1982)  
 M. Louise Curley (1982)  
 Robert Eisner (1982)  
 Irma Adelman (1983)  
 Janet C. Goulet (1983)  
 Jean A. Shackelford (1983)  
 Monique Garrity (1984)  
 Joan G. Haworth (1984)  
 Nancy D. Ruggles (1984)  
 Gail Wilensky (1984)  
 Gardner Ackley, *ex officio*

*Committee on U.S.-Soviet Exchanges*

Lloyd G. Reynolds, *Chair* (1982)  
 Abram Bergson (1982)  
 Joseph A. Pechman (1982)  
 Richard N. Rosett (1982)  
 Rendigs Fels, *ex officio*

## COUNCIL AND OTHER REPRESENTATIVES

*American Association for the Advancement of Science Section K on Social, Economic, and Political Sciences*

Roger Bolton (1983)

*American Association for the Advancement of Slavic Studies*

Elizabeth Clayton (1982)

*AEA/SSRC-Joint Committee on U.S.-China Exchanges*

Gregory Chow, *AEA Chair* (1983)  
 Kenneth J. Arrow  
 Lawrence R. Klein  
 Theodore W. Schultz

*American Council of Learned Societies*

C. Elton Hinshaw (1986)

*Consortium of Social Science Associations (COSSA)*

Henry J. Aaron  
Joseph A. Pechman

*Federal Statistics Users Conference*

Paul Wonnacott (1982)

*International Economic Association*

Anne O. Krueger (1982)  
C. Elton Hinshaw (1985)

*Policy Board of the Journal of Consumer Research*

Lester Telser (1982)

*National Archives Advisory Council—General Services Administration*

William N. Parker (1984)

*National Bureau of Economic Research*

Carl F. Christ (1984)

*Seventh Symposium on Statistics and the Environment—Steering Committee*

Eugene Seskin  
Paul Portney

*Social Science Research Council*

Hugh Patrick (1984)

*SSRC—Committee of Professional Associations on Federal Statistics (COPAFS)*

John H. Cumberland (1982)  
Gary Fromm (1983)

*U.S. National Commission for UNESCO*

Walter S. Salant

## REPRESENTATIVES OF THE ASSOCIATION ON VARIOUS OCCASIONS—1982

*Inaugurations*

Bernard W. Harleston, The City College of  
The City University of New York

Leo Troy

William J. Teague, Abilene Christian University

Mary S. Staig

Paula P. Brownlee, Holling College

Bruce Herrick

J. Russell Nelson, Arizona State University

Leahmae McCoy

Robert L. Randolph, Alabama State University

Samuel G. Thangiah

Betty L. Siegel, Kennesaw College

Jo Ann Jones

Robert F. Sasseen, The University of Dallas

G. Gardner Williams

Curtis L. McCray, The University of North Florida

Arthur R. Dorsch

Alfred F. Hurley, North Texas State University & Texas College of Osteopathic Medicine

Samuel Bostaph

Barry B. Thompson, Tarleton State University

James A. Myers

John M. Howell, East Carolina University

Robert M. Fearn

## ASSA 1982 CONVENTION COMMITTEE

Peter Fousek, *Chair*

Alice Christensen, *Vice Chair*

Barbara Weaver, *Convention Manager*

Jean-Ellen Giblin

Corine Parodi

Noralynn Marshall

Leslie Fleishman

Arthur W. Samansky

Violet O. Kohsman

Marilyn Rubin

Catherine Kweit

Janet Aschenbrenner

Marlene Hall

Norma Ayres

Lois Banks

C. ELTON HINSHAW, *Secretary*

# Report of the Treasurer For the Year Ending December 31, 1982

The Executive Committee of the American Economic Association is continuing the policy of increasing dues less than the general rate of inflation, thus reducing dues in real terms. Between January 1, 1976, and January 1, 1982, dues increased 24 percent compared to a rise in the implicit consumption deflator of 55 percent, implying a decline in real dues on the order of 20 percent. With inflation now running at an annual rate of 5 percent, the 3 percent increase in dues effective January 1, 1983, means a small further dues reduction in real terms.

Continuation of this policy has been made possible by raising the price of subscriptions

to nonmembers. When audited figures become available, a substantial surplus is expected for 1982 and a smaller surplus in 1983 in spite of rapid increases in expenses. As Table 1 shows, proposed expenditures for 1983 are more than 10 percent higher than the amount budgeted for 1982 and 25 percent higher than the actual expenditures of 1981. Expected revenues for 1983 are only 8.5 percent higher than 1981. If continued, these trends would result in a deficit in 1984. But for the moment, the finances of the Association are robust.

RENDIGS FELS, *Treasurer*

TABLE 1—AMERICAN ECONOMIC ASSOCIATION BUDGETS, 1982–83  
(Thousands of dollars)

	First Nine Months, Actual (Unaudited)		Full Year		
	1981	1982	Actual 1981	Budgeted	
				1982	1983 <sup>a</sup>
<b>REVENUES FROM DUES AND ACTIVITIES</b>					
Membership dues	506	540	682		
Nonmember subscriptions	451	479	621		
Subtotal	958	1019	1303	1300 <sup>c</sup>	1385
JOE subscriptions	14	16	22	22	30
Advertising	69	62	99	100	95
Sales of <i>Index of Economic Articles</i>	49	43	66	50	50
Sales of copies, etc.	26	26	34	33	35
Sale of mailing list	22	23	39	35	40
Annual meeting	14	37	14	15	15
Sundry	21	34	51	28	40
Total, dues and activities	1173	1260	1626	1583	1690
<b>INVESTMENT GAINS</b>					
Net revenues	15	104	110	140	170
	1188	1364	1737	1723	1860
<b>PUBLICATION EXPENSES</b>					
<i>American Economic Review</i>	316	349	407	450	493
<i>Journal of Economic Literature</i>	423	458	564	631	702
<i>Directory Publication</i>	49	41	65	70	75
<i>Job Openings for Economists</i>	26	27	42	44	50
<i>Index of Economic Articles</i>	20	19	27	30	30
Subtotal	833	895	1105	1225	1350
<b>OPERATING AND ADMINISTRATIVE EXPENSES</b>					
Extraordinary items				48 <sup>b</sup>	35 <sup>d</sup>
General and administrative	181	240	257	267	287
Committees	22	26	39	43	50
Federal income taxes	7	4	5	12	5
Subtotal	211	270	301	370 <sup>c</sup>	377
Total expenses	1043	1165	1406	1595	1727
<b>REVENUES IN EXCESS OF EXPENSES</b>					
	144	199	331	128	133

<sup>a</sup>As approved by the Executive Committee on December 27, 1982.

<sup>b</sup>Includes \$35,000 for COSSA, \$10,000 for the Economics Institute, and \$3,000 for COPAFS.

<sup>c</sup>\$3,000 higher than budget published in *American Economic Review Proceedings*, May 1982, p. 411, as a result of special appropriation for COPAFS.

<sup>d</sup>For COSSA.

## Report of the Finance Committee\*

The accompanying inventory summary lists the securities held by the American Economic Association as of December 31, 1982, with costs and market values as of that date. The total market value of the securities portfolio at year end was \$2,806,458. After making adjustments for cash additions and withdrawals, we estimate that the Association's investment portfolio experienced a total investment return of +27.5 percent during 1982. The equities held in the portfolio continued their past record of attractive performance generating a total return of +28.8 percent. This return compares favorably with the Standard and Poor's 500 average (+21.4 percent) and other market averages.

Economic conditions as well as the environment for the securities markets shifted rather dramatically during the year. In anticipation of these evolving conditions, several portfolio changes were made last year. These involved new commitments of individual issues held at year end in Pfizer, American Greetings Corp., Citicorp, Commerce Clearing House, Schlumberger Ltd., Wang Labs, Datapoint, Denny's, Baxter Travenol Labs, Emerson Electric, and R. H. Macy & Co., and the elimination of individual holdings of

Hughes Tool Co., Lear Siegler, Litton Industries, Celeron Corp., Tidewater Inc., Ocean Drilling & Explor., Smithkline Ltd., Halliburton, and Texas Commerce Bank. Fixed-income securities continued to have a medium-term orientation with an average weighted maturity of 5.6 years.

Nineteen eighty-two was a year in which the cumulative effect of much lower inflation and weak economic conditions were finally reflected in sharply falling interest rates beginning at midyear. This turn of events sparked a strong rally in both the bond and stock markets. As rates continued to decline throughout the fall, this rally was sustained; and, as mentioned above, the impact on the Association's portfolio was quite favorable.

Looking into 1983, investor interest appears focused on two questions: the direction of economic activity and the government's monetary and fiscal policy. It is the expectation of the fund's investment advisor that the coming year will see a recovery of economic activity although at a rate below the average for past recovery periods. This should allow for the maintenance of lower inflation rates and meaningful improvement in corporate profitability later in the year. While, at this writing, it is impossible to describe the nature of a fiscal policy compromise, it is also assumed that political and economic conditions will push the administration and Congress to develop a program of budget cuts (including some in defense), expenditure stretchouts, and revenue increases leading to downward trending deficits in fiscal year

\*The Report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some modest discrepancies between figures in the Report of the Finance Committee and the Auditors' Report which will be published in a later issue of the *Review*.

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 31, 1982

	Value	Percent	Estimated Income	Estimated Current Yield
Cash Equivalents	83,766	3.0	7,120	8.5
Short-Term Securities	332,627	11.9	46,313	13.9
Medium-Term Securities	271,186	9.7	34,138	12.6
Long-Term Securities and Preferred Stocks	50,379	1.8	5,625	11.2
Convertible Securities	74,050	2.6	8,726	11.8
Equity Securities	1,994,450	71.1	58,678	2.9
Total	2,806,458	100.0	160,600	5.7

1984 and beyond. Given the historically low relative level of stock prices (despite the strikingly strong 1982 rally) as well as expectations for a further period of lower inflation and interest rates, the environment for common stocks over the next twelve to eighteen months looks positive.

Keeping these factors in mind as well as recognizing the Association's favorable financial position, the Finance Committee decided at its meeting in December to leave unchanged its current directives to Stein Roe

& Farnham, our investment advisor. These directives provide that 50 to 75 percent of the value of the portfolio should be invested in equities except that, if a rise in the stock market carries the proportion above 75 percent, the investment counsel is not required to sell stocks. They further provide that the average maturity of fixed-dollar assets shall not exceed eight years.

RENDIGS FELS, *Chair*

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 31, 1982

	Amount	Price	Value	Unit Cost	Total Cost	Estimated Income
<b>Cash Equivalents and Fixed-Income Securities (26.3 percent)</b>						
<i>Cash Equivalents (0–1 year)(11.4 percent)</i>						
Stein Roe Cash Reserves, Inc.	83,765	1	83,766	1	83,768 <sup>a</sup>	7,120
Subtotal Cash Equivalents (0–1 year)			83,766		83,768	7,120
<i>Other Short-Term Securities (1–5 years)(45.1 percent)</i>						
Fed Home Loan Bks (15.800 01/25/84)	50,000	107	53,344	100	50,000	7,900
Fed Natl Mtg Assn (17.200 11/12/84)	50,000	112	56,094	100	50,000	8,600
Fed Home Loan Bks (13.900 07/25/85)	50,000	109	54,250	100	50,000	6,950
Fed Farm Cr Bks (15.800 01/20/86)	50,000	114	56,813	100	50,000	7,900
Fed Home Loan Bks (15.300 02/25/86)	50,000	113	56,313	100	50,000	7,650
Fed Natl Mtg Assn (14.625 06/10/86)	50,000	112	55,813	100	49,953	7,313
	300,000		332,627		299,953	46,313
Subtotal Other Short-Term Securities (1–5 years)			332,627		299,953	46,313
<i>Medium-Term Securities (5–10 years)(36.8 percent)</i>						
Fed Home Loan Bks (15.100 02/27/89)	50,000	117	58,500	100	50,000	7,550
Crdthrift Finl Nt (10.250 04/15/89)	50,000	95	47,256	84	42,039	5,125
US Treas Nts (14.500 07/15/89)	50,000	117	58,375	99	49,733	7,250
Florida Pwr 1st (13.300 11/01/90)	50,000	104	51,962	95	47,598	6,650
Duke Pwr 1st Mtg (15.125 03/01/91)	50,000	110	55,093	99	49,625	7,563
	250,000		271,186		238,995	34,138
Subtotal Medium-Term Securities (5–10 years)			271,186		238,995	34,138
<i>Long-Term Securities (More than 10 years)(6.8 percent)</i>						
Hydro-Quebec Debentures (11.250 10/15/09)	50,000	101	50,379	96	48,031 <sup>a</sup>	5,625
Subtotal Long-Term Securities			50,379		48,031	5,625
Total Cash and Fixed-Income Securities			737,958		670,747	93,196
<b>Convertible Securities (Limited Risk)(2.6 percent)</b>						
<i>Convertible Bonds (100.0 percent)</i>						
Datapoint CV (8.875 06/01/06)	50,000	66	32,750	67	33,250	4,438
Barnett Banks CV (12.250 12/15/06)	35,000	118	41,300	100	35,000	4,288
			74,050		68,250	8,726
Total Convertible Securities (Limited Risk)			74,050		68,250	8,726
<b>Equity Securities (71.1 percent)</b>						
<i>Energy Services (2.1 percent)</i>						
Schlumberger Ltd	900	47	41,963	31	28,238	864
<i>Food, Beverages, and Tobacco (5.2 percent)</i>						
General Cinema	1,000	29	29,125	16	16,215	520
Genl Cinema \$.64 CV PFD	1,000	27	26,500	17	16,635	640
Philip Morris	800	60	48,000	22	17,726	1,920
			103,625		50,576	3,080

TABLE 2—(Continued)

	Amount	Price	Value	Unit Cost	Total Cost	Estimated Income
<i>Publishing (3.8 percent)</i>						
Amer Greetings Corp CL A	1,000	38	37,750	26	26,125	680
Commerce Clearing House	600	62	37,200	49	29,100	936
			74,950		55,225	1,616
<i>Drugs and Hospital Supplies (8.7 percent)</i>						
Abbott Lab	1,500	39	58,125	11	16,943 <sup>a</sup>	1,260
Baxter Travenol Labs	800	48	38,700	49	38,800	368
Merck	500	85	42,313	57	28,402 <sup>a</sup>	1,400
Pfizer	500	69	34,438	56	27,880	1,160
			173,576		112,025	4,188
<i>Petroleum (3.0 percent)</i>						
Atlantic Richfield	1,000	42	42,000	36	35,823	2,400
Standard Oil Ohio	510	36	18,169	20	10,076 <sup>a</sup>	1,326
			60,169		45,899	3,726
<i>Machinery; Fabr Metal Product (2.6 percent)</i>						
Diebold Inc	700	75	52,150	49	34,501 <sup>a</sup>	700
<i>Computers and Office Equipment (4.7 percent)</i>						
IBM	480	96	46,201	28	13,325 <sup>a</sup>	1,651
Wang Labs CV (10.000 11/15/06)	30,000	157	47,100	95	28,500	3,000
			93,301		41,825	4,651
<i>Electrical Equipment (5.4 percent)</i>						
Emerson Electric	700	61	42,350	61	43,036	1,470
General Electric	690	95	65,465	36	24,536 <sup>a</sup>	2,346
			107,815		67,572	3,816
<i>Photography (2.2 percent)</i>						
Eastman Kodak	500	86	43,000	47	23,740	1,775
<i>Telephone Companies (3.3 percent)</i>						
MCI Communications	1,800	36	65,475	17	31,388	
<i>Broadcasting &amp; Communications (3.0 percent)</i>						
Metromedia	200	296	59,200	142	28,498 <sup>a</sup>	1,400
<i>Electric Utilities (4.1 percent)</i>						
Houston Inds	1,600	20	32,000	19	30,272	3,456
Texas Utilities	2,100	24	49,350	22	46,431	4,284
			81,350		76,703	7,740
<i>Sanitary Services (3.2 percent)</i>						
Waste Management	1,200	54	64,200	30	35,646	624
<i>Distribution (11.3 percent)</i>						
Macy R H & Co Inc	600	61	36,600	64	38,421	600
McDonalds	900	60	54,338	40	36,315	792
Wal-Mart Stores	2,000	50	99,750	18	35,350	360
Dennys Conv Debentures (9.500 10/15/07)	30,000	112	33,600	100	30,000	2,850
			224,288		140,086	4,602
<i>Banks; Savings and Loans (1.6 percent)</i>						
Citicorp	1,000	33	32,500	27	26,595	1,720
<i>Services (7.8 percent)</i>						
Humana Inc	2,000	45	90,000	12	24,962 <sup>a</sup>	1,600
Warner Communications	1,000	34	33,500	24	24,232	1,000
Graphic Scan CV (10.000 12/01/01)	30,000	109	32,775	100	30,000	3,000
			156,275		79,194	5,600
<i>Miscellaneous (28.1 percent)</i>						
Stein Roe Special Fund	4,174	14	58,823	12	49,428 <sup>a</sup>	1,211
Stein Roe Universe Fund Inc	7,892	64	501,790	51	398,755 <sup>a</sup>	11,365
			560,613		448,183	12,576
TOTAL EQUITY SECURITIES			1,994,450		1,325,894	58,678
TOTAL SECURITIES AND CASH			2,806,458		2,064,891	160,600

<sup>a</sup> More than one cost basis.

# Report of the Managing Editor

## *American Economic Review*

My second year year as managing editor of the *Review* has been no less busy than the first, but greater familiarity with the work—plus able help from John Riley (the new associate editor) in manning the barricades—has made the burdens of the job seem a good deal lighter.

There have been no conscious changes in editorial procedures or policies as compared with last year. We are continuing to dispense with outside editorial scanners in order to avoid delays in processing manuscripts. We are also continuing to send comments on previously published papers to the original authors for their information and reaction before seeking an independent review from members of the Board of Editors or other disinterested parties. This procedure is less than perfect, but it seems to be working well enough to be retained until a better alternative presents itself. We are also continuing to make every effort to ensure that, except in most unusual circumstances, authors have a final editorial decision within three months of the date that their manuscript is received at the UCLA editorial office (some idea of the success of this effort may be gleaned from the data presented in Tables 4 and 5 below).

### Operations

The recent history of manuscript submissions and papers published is shown in Ta-

bles 1 and 2. We received 820 papers in 1982. Like the figure for 1981, this is well above the annual average for the years 1975–80, but it is not significantly different from last year. Comparison of monthly rates of manuscript submission for 1981 and 1982 suggests that the present annual level of manuscript flow is likely to continue for the immediate future.

The disposition of manuscripts received during 1981 and 1982 is shown in Table 3.

TABLE 1—MANUSCRIPTS SUBMITTED  
AND PUBLISHED, 1963–82

Year	Submitted	Published	Ratio of Published-to-Submitted
1963	329	46	.14
1964	431	67	.16
1965	420	59	.14
1966	451	62	.14
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15

TABLE 2—SUMMARY OF CONTENTS, 1981 AND 1982

	1980		1982	
	Number	Pages	Number	Pages
Articles	53	711	52	761
Shorter Papers, including Comments and Replies	62	367	68	407
Dissertations		12		21
Announcements and Notes Section		55		52
Index		11		10
Total	115	1156	120	1251

TABLE 3—DISPOSITION OF MANUSCRIPTS, 1981 AND 1982

	1981	1982
Manuscripts Received	784	820
Completed Processing:	690	694
Accepted	42	113
Rejected	648	581
Acceptance Rate	6.1%	13.8%
Currently in Process	94	126

TABLE 4—DISTRIBUTION OF EDITORIAL DECISION LAGS BETWEEN RECEIPT AND REJECTION, JANUARY 1–OCTOBER 30, 1982

Weeks to Rejection	Number of Manuscripts	Percent
0–4	244	.51
5–9	137	.28
10–14	54	.11
15–19	23	.05
20–24	13	.03
25–29	4	.01
30+	7	.01
Total	482	100.

The acceptance rate for 1982 is roughly the same as the ratio of published-to-submitted papers shown in Table 1 (15 percent). The rise in this rate as compared with last year reflects a return to normality (the unusually low rate for 1981 was a consequence of efforts to reduce an outstanding backlog of papers awaiting publication). The file of accepted papers as of December 31, 1982, contained 58 manuscripts; 26 of these will appear in March 1982 and most of the remainder will be published in June.

Additional information about editorial office processing lags is provided in Tables 4 and 5. The average inventory of manuscripts "in process" has increased substantially during the past year. This appears to be a consequence of increased time lags between receipt and rejection, which derives from an increase in the proportion of manuscripts that are subjected to external review. A fair number of papers are still being rejected on the basis of internal screening, but the number of these has dropped substantially as compared with last year.

TABLE 5—AVERAGE PUBLICATION LAGS, BY JOURNAL ISSUE

Issue	Number of Weeks Lag		
	Receipt to Acceptance	Acceptance to Publication	Receipt to Publication
September 1982	15	58	73
December 1982	20	40	60
March 1983	24	41	65

TABLE 6—SUBJECT MATTER DISTRIBUTION OF PUBLISHED MANUSCRIPTS, 1981 AND 1982

	Published	
	1981	1982
General Economics and General Equilibrium Theory	6	6
Microeconomic Theory	20	17
Macroeconomic Theory	4	4
Welfare Theory and Social Choice	11	8
Economic History, History of Thought, Methodology	7	3
Economic Systems	1	2
Economic Growth, Development, Planning, Fluctuations	3	13
Economic Statistics and Quantitative Methods	6	6
Monetary and Financial Theory and Institutions	8	11
Fiscal Policy and Public Finance	6	8
International Economics	13	8
Administration, Business Finance	5	2
Industrial Organization	10	7
Agriculture, Natural Resources	1	3
Manpower, Labor Population	10	14
Welfare Programs, Consumer Economics, Urban and Regional Economics	4	8
Total	115	120

TABLE 7—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1982 AER

	Copies Printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	28,000	288	328	\$51,506.28	\$1,537.51	\$53,043.79
May	28,000	442	472	67,797.55	2,525.16	70,322.71
June	27,500	332	352	51,878.45	1,762.51	53,640.96
September	27,500	302	344	50,733.12	1,098.01	51,831.13
December <sup>b</sup>	27,500	319	376	54,407.57	2,000.00	56,407.57
Annual Misc. <sup>a</sup>						10,000.00
Total		1,683	1,872	\$276,322.97	\$8,923.19	\$295,246.16

<sup>a</sup>Estimated: based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

<sup>b</sup>Estimated.

The subject matter distribution of pages published in 1981 and 1982 is shown in Table 6. As indicated, some changes have occurred but none appears to be significant. Popular opinion to the contrary notwithstanding, what appears in the *Review* depends much more upon what is submitted than upon the idiosyncracies of the managing editor.

#### Expenses: Printing and Mailing

Table 7 shows the printing and mailing expenses for the four regular issues and for the *Papers and Proceedings* issue of the *Review* for 1982. As in earlier years, the *Papers and Proceedings* accounted for approximately 25 percent of total printing and mailing expenses. Total printing and mailing expenses rose about 11 percent over last year (from roughly \$270,000 in 1981 to \$295,000 in 1982). The increase in costs for the coming year is expected to be considerably less (approximately 5 percent), but this is on the assumption that costs over which we have no control (mailing and postage) do not change appreciably.

#### Papers and Proceedings

The fifth volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in 1982. This task was ably handled by Wilma St. John and Theresa De Maria, with modest assistance from the managing editor.

The *Proceedings* volume continues to impose a heavier burden on the editorial staff

than is indicated by its cost relative to regular issues of the *Review*. Last year I voiced the hope that, with the office move completed, the production of the *Proceedings* volume would involve fewer difficulties, but that hope has not been realized. The truth seems to be that there is no way to fit the publication of the *Proceedings* volume into our regular work schedule without creating serious temporary overloads. In an attempt to resolve some of these problems, we plan to go directly from manuscript to page proofs in the 1983 *Proceedings* volume. I shall report on the outcome of this experiment next year.

#### Board of Editors

The Board of Editors presently consists of eighteen members, chosen by the managing editor, with the approval of the Executive Committee of the Association. The Board is particularly helpful in dealing with comments on published articles and in reading papers that are the subject of complaint over the fairness of competence of referees. Members of the Board also advise the managing editor on editorial policy and occasionally serve as referees of particularly difficult manuscripts. Needless to say, I am grateful to members of the Board for their assistance and advice, for their moral support, and for occasional snippets of information about authors' reactions to letters of rejection and acceptance from the managing editor.

I plan to make some changes during the next few years in the membership and duties of the Board. Specifically, I endorse the

suggestion of the AEA *Ad Hoc* Committee on Publications that the Board of Editors of the *Review* be substantially reduced in size.

Six members of the present Board will complete their terms on March 31, 1983: Albert Ando, Herschel Grossman, Peter Howitt, Anne Krueger, James Smith, and Robert Willig. I thank them for work well done, and I wish them an easier life in the years ahead. I also thank the continuing members of the Board for services performed and in prospect: Gerald Bierwag, Thomas Cooley, Ronald Ehrenberg, Ted Frech, Larry Kotlikoff, Roy Weintraub, George Akerlof, Patricia Danzon, Jack Hirshleifer, Rick

Mishkin, Sherwin Rosen, and Richard Schmalensee.

### Acknowledgments

I wish to thank my office associates, Wilma St. John, Lois Bagley, and Theresa De Maria for dedicated performance that is frequently beyond the call of duty. Warm thanks are also due to our graduate mathematics consultant, Peter Swank. Finally, I want to thank and express my admiration for the more than 550 referees whose efforts, though largely unsung, have so crucially influenced virtually every paper that appears in the *Review*.

M. Abbott	D. Benjamin	J. M. Buchanan	R. Craine
A. B. Abel	Y. Ben-Porath	C. Bull	J. T. Cuddington
M. Adler	G. Benston	J. Bulow	A. Cukierman
R. S. Ahlbrandt	E. Berglas	E. Burmeister	G. C. Daly
D. J. Aigner	T. C. Bergstrom	P. G. Burrows	S. H. Danziger
R. Aiyagari	E. R. Berndt	M. Burstein	P. Danzon
G. Akerlof	R. A. Berry	G. T. Burtless	M. Darby
A. Alchian	J. Bhagwati	G. Butters	S. P. Das
A. Aldo	G. O. Bierwag	W. Butz	P. S. Dasgupta
R. J. Allard	F. Black	P. Cagan	P. Davidson
J. G. Altonji	C. Blackorby	G. Calvo	J. B. Davies
E. Ames	O. J. Blanchard	C. Campbell	R. Day
J. E. Anderson	A. S. Blinder	D. R. Capozza	R. Deacon
A. K. Ando	G. C. Blomquist	J. A. Carlson	H. De Angelo
C. Ansley	B. Bluestone	D. Carlton	A. V. Deardorff
M. Aoki	R. W. Boadway	J. Carr	A. S. Deaton
R. Arnott	Z. Bodie	C. A. Carter	H. Demsetz
W. B. Arthur	L. A. Boland	A. Charnes	D. N. DeTray
A. J. Auerbach	T. E. Borchering	A. C. Chiang	A. De Vany
R. D. Auerbach	G. Borts	P. Chinloy	D. Dewey
C. Azariadis	M. J. Boskin	J. S. Chipman	P. A. Diamond
R. Bade	K. Boyer	C. Christ	W. E. Diewert
M. J. Bailey	R. S. Boyer	J. Ciccolo	W. Dolde
M. N. Baily	R. Braeutigam	K. Clark	R. Dornbusch
R. E. Baldwin	M. Bray	D. L. Cleeton	M. Dotsey
P. Bardhan	F. P. Brechling	C. T. Clotfelter	A. Drazen
A. J. Barkume	H. Brems	R. M. Coen	J. H. Dreze
D. P. Baron	T. F. Bresnahan	R. A. Cohn	J. Eaton
J. M. Barron	D. L. Brito	W. E. Conrad	B. Eden
Y. Barzel	H. W. Brock	G. M. Constantinides	S. Edwards
R. W. Batchelder	M. Bronfenbrenner	T. F. Cooley	R. G. Ehrenberg
W. Baumol	C. C. Brown	L. W. Copithorne	I. Ehrlich
G. S. Becker	J. Brown	P. N. Courant	R. Eisner
S. Beckett	E. K. Browning	G. A. Craig	B. C. Ellickson
L. Benham	J. Bryant	M. Crain	C. M. Engel

S. Engerman	D. Goldsbrough	L. Hurwicz	S. A. Lippman
M. P. Engers	A. C. Goodman	F. O. Irvine, Jr.	S. C. Littlechild
D. Epple	M. J. Gordon	R. M. Isaac	R. Lombra
T. W. Epps	R. J. Gordon	K. Iwai	G. C. Loury
W. Ethier	M. Gort	L. F. Jackson	M. C. Lovell
K. Evans	J. P. Gould, Jr.	D. Jaffee	R. E. Lucas, Jr.
O. Evans	D. A. Graham	J. Johnston	M. Lurie
R. C. Fair	D. Graham	R. Jones	S. Lustgarten
R. Faith	P. Graves	L. Jonung	R. A. McCain
R. E. Falvey	K. V. Greene	P. L. Joskow	T. S. McCaleb
E. F. Fama	R. G. Gregory	R. Just	J. J. McCall
L.-S. Fan	G. M. Grossman	J. P. Kalt	B. T. McCallum
H. S. Farber	H. I. Grossman	D. P. Kamerschen	D. McCloskey
D. J. Faurot	H. G. Grubel	M. Kamien	H. McCulloch
E. L. Feige	A. Guha	R. Kanbur	I. M. McDonald
A. M. Feldman	M. Hadjimichalakis	E. Katz	R. L. McDonald
M. Feldstein	R. Hall	J. B. Kau	J. M. McDowell
S. Ferris	J. Haltiwanger	M. C. Keeley	W. McEachern
R. Findlay	D. Hamberg	M. S. Khan	D. McFadden
J. M. Finger	D. S. Hamermesh	R. Kihlstrom	G. W. McKenzie
S. Fischer	W. L. Hansen	B. Klein	R. I. McKinnon
P. C. Fishburn	H. B. Hansmann	M. Kohn	C. D. Macrae
F. M. Fisher	J. A. Hanson	R. H. Koller II	G. S. Maddala
M. Fisher	E. A. Hanushek	R. W. Kopcke	R. R. Maddock
R. P. Flood	J. Harkness	R. Kormendi	J. H. Makin
L. R. Forest, Jr.	L. Harris	M. Kosters	L. Makowski
E. M. Foster	A. Harrison	L. Kotlikoff	E. Mansfield
M. Fraenkel	D. Harrison	M. B. Krauss	J. Markusen
J. Frankel	G. Harrison	A. Krueger	J. Marshall
H. E. Frech III	O. Hart	P. Krugman	E. Maskin
A. M. Freeman	D. Hartman	V. Kwok	T. Mayer
R. Freeman	M. Hashimoto	J. E. Kwoka	D. Mayers
K. R. French	J. A. Hausman	J.-J. Laffont	J. Mayshar
J. A. Frenkel	R. H. Haverman	K. J. Lancaster	J. L. Medoff
J. S. Fried	G. Hay	E. Leamer	M. Melvin
B. M. Friedman	Y. Hayami	N. Leff	C. F. Menezes
D. Friedman	F. Hayashi	K. Leffler	R. A. Meyer
D. D. Friedman	J. J. Heckman	A. Leibowitz	W. H. Miernyk
J. Friedman	R. Heiner	A. Leijonhufvud	P. Milgrom
M. Friedman	M. F. Hellwig	H. Leland	R. Miller
R. Frydman	E. Helpman	S. F. LeRoy	T. C. Mills
D. Fudenberg	W. Hirsch	M. Levi	H. P. Minsky
E. G. Furubotn	J. Hirshleifer	H. M. Levin	F. Mishkin
C. A. Futia	R. J. Hodrick	R. C. Levin	E. J. Mitchell
G. Galles	C. A. Holt	D. Levine	R. Moffitt
F. Gallop	D. Holthausen	H. G. Lewis	J. C. Moore
P. M. Garber	B. Hool	W. A. Lewis	D. Mortensen
J. Geweke	P. M. Horvitz	G. D. Libecap	J. Muellbauer
R. J. Gilbert	J. R. Hosek	C. Lieberman	D. Mueller
C. Gilroy	P. Howitt	S. J. Liebowitz	T. Muench
L. Girton	E. A. Hudson	D. M. Lilien	D. J. Mullineaux
F. Glahe	C. R. Hulten	C. Lim	Y. Mundlak
V. P. Goldberg	A. P. Hurter	C. M. Lindsay	M. Mussa

R. F. Muth	R. Rasche	J. B. Shoven	H. Varian
T. Negishi	A. Raviv	J. J. Siegel	R. Verrecchia
G. R. Neumann	A. Razin	E. Silberberg	W. K. Viscusi
D. M. G. Newbery	C. W. Reimers	L. Simon	I. T. Vogelsang
G. R. Newman	J. Reinganum	C. A. Sims	G. von Furstenberg
P. K. Newman	W. J. Rieber	L. A. Sjaastad	C. C. von Weizsacker
A. L. Nichols	J. Riley	J. Skinner	M. Wachter
D. A. Nichols	A. L. Robb	K. A. Small	D. A. Walker
S. J. Nickell	M. Robinson	J. Smith	N. Wallace
D. C. North	A. J. Robson	L. B. Smith	M. Ward
W. E. Oates	H. Rockoff	V. L. Smith	R. Weber
M. Obstfeld	J. Roemer	E. Smolensky	R. Weintraub
G. O'Driscoll	R. W. Roll	J. Sobel	A. Weiss
M. Okuno	T. Romer	E. Solomon	L. Weiss
M. Olson, Jr.	S. Rose-Ackerman	G. Solon	L. W. Weiss
J. Ostroy	K. T. Rosen	R. M. Solow	Y. Weiss
M. Ott	S. Rosen	H. Somers	M. Weitzman
M. Paglin	S. Ross	C. S. Spatt	F. Welch
J. C. Panzar	J. J. Rotemberg	J. Spraos	K. L. Wertz
M. Parkin	J. Rowley	D. F. Spulber	D. N. Westcott
R. W. Parks	D. L. Rubinfeld	R. Startz	J. F. Weston
J. Paroush	R. Ruffin	M. Staten	L. E. Westphal
C. J. Parsley	J. Sachs	J. L. Stein	J. Weymark
D. K. Pearce	S. W. Salant	J. E. Stiglitz	J. K. Whitaker
J. Peek	J. Salmon	A. C. Stockman	B. White
S. Pejovic	P. A. Samuelson	M. Straszheim	L. H. White
J. H. Pencaval	W. F. Samuelson	C. Stuart	J. A. Wilcox
M. K. Perry	A. Sandmo	L. H. Summers	L. L. Wilde
B. Pesek	A. M. Santomero	J. Sutton	S. Wildman
D. Phaup	D. Sappington	J. Svejnar	S. Williamson
C. Phelps	T. J. Sargent	P. Swan	R. D. Willig
E. S. Phelps	F. M. Scherer	V. Tanzi	R. J. Willis
D. A. Pierce	R. Schmalensee	J. A. Tatom	C. A. Wilson
R. S. Pindyck	P. Schmidt	J. B. Taylor	R. Wilson
J. E. Pippenger	M. D. Schmitz	D. Teece	G. C. Winston
M. Plant	M. S. Scholes	L. G. Telser	S. G. Winter
R. D. Plotnick	A. J. Schwartz	P. Temin	A. D. Witte
C. Plourde	A. R. Schwartz	G. Thomas	A. M. Wojnilower
D. Poirer	M. Schwartz	E. Thompson	P. Wonnacott
S. Polachek	J. Seater	T. N. Tideman	R. J. Wonnacott
R. A. Pollak	L. Seidman	J. Tirole	D. O. Wood
H. O. Pollakowski	T. Sekine	J. Tobin	S. Woodward
W. Poole	A. K. Sen	M. P. Todaro	L. B. Yeager
R. H. Porter	W. Shaffer	N. Tomes	J. Yinger
J. M. Poterba	C. Shapiro	R. Topel	H. Yoshikawa
E. C. Prescott	P. Shapiro	S. Turnovsky	L. Young
F. L. Pryor	L. Shapley	L. Tyson	B. Yu
G. Pyatt	S. Shavell	B. D. Udis	E. E. Zajac
J. Quinn	R. Sherman	J. Umbeck	V. Zarnowitz
J. Quirk	R. J. Shiller	J. Vanderkamp	R. Zeckhauser
G. Ranis	Y. Shilony	R. A. Van Order	G. R. Zodrow

## Report of the Managing Editor

### *Journal of Economic Literature*

The *Journal* continued to appear during 1982, operating under the general plans and procedures established in 1981. The documentation services of the *Journal* are carried on in Pittsburgh under the leadership of the *Journal's* senior associate editor, Naomi Perlman. Drucilla Ekwurzel continues as assistant editor in the Pittsburgh office. The documentation services comprise five departments of the *Journal*, that is, New Books: An Annotated Listing; Contents of Current Periodicals; Subject Index of Articles in Current Periodicals; Selected Abstracts; and Index of Authors of Articles in the Subject Index. The annual *Index of Economic Articles* is also prepared and edited in Pittsburgh. The Articles and Book Review Departments of the *Journal* are edited at the Stanford office. Alexander Field serves as associate editor in charge of book reviews. John Pencavel became an associate editor in September 1982. He shares responsibility with the managing editor for the articles. Anne Rawley Saldich continues as assistant editor in charge of production and staff at Stanford.

During 1982, the *Journal* published seventeen titles, including nine articles, three review articles, and five substantial communications. There were 160 book reviews. The volume of documentation continues to expand slowly. Some 1,325 books were listed together with annotations. They are classified by subject matter. The preparation of the annotations, which demands much skill and care, is a substantial burden which we think is justified by their value to readers. Over the course of the year, the *Journal* classified and indexed nearly 8,500 articles appearing in 1,047 issues of 262 journals published in the United States, Canada, and many other countries. The *Journal* also carried 1,665 abstracts of articles.

The annual *Index of Economic Articles* for 1978, which provides subject and author indexes of articles appearing in that year, is

scheduled for publication in the spring of 1983. It includes not only the articles originally indexed in the quarterly issues, but also those published in some 200 to 300 collective volumes—conference proceedings, *Festschriften*, and other occasional volumes. I should like to acknowledge the very valuable contribution of Asatoshi Maeshiro of the University of Pittsburgh who is responsible for much of the work of classification which lies behind the indexing of books and articles in both the *Journal* and annual *Index*.

The Pittsburgh office of *JEL* has arranged for Dialog Information Retrieval Service to provide online computer access (in early 1983) to the *JEL* and *Index* data bases, annotations, and abstracts. The data base file contains complete bibliographic citations to articles indexed in the *Index of Economic Articles* (1969–77) and quarterly *JEL* issues through 1982. The file also includes articles in collective volumes (200–300 books per year) included in the 1969–77 *Indexes* and government documents listed in the 1969–72 *Indexes*. This file may be searched using author, journal, subject index title, geographic area, date, and other search strategies, including the *Index* four-digit classification number.

Dialog will soon have available *JEL* abstracts and annotations, beginning with the June and September 1982 issues, respectively. In the future, Dialog will add tapes from the quarterly *JEL* issues and the annual *Index* soon after they are prepared for publication.

The editorial objectives and policies of the *Journal* were set forth in an Editor's Note in the June 1981 issue, p. 491. In accordance with this statement, the managing editor commissions the *Journal's* expository, survey and review articles. The *Journal*, however, welcomes proposals for such articles.

The Association owes a warm debt of gratitude to the 1982 Board of Editors and Consulting Editors who helped plan and re-

view the *Journal's* articles, and to the many economists who served as referees. I should like to notice with special thanks the contributions of Nicholas Balabkins who now completes a term as a member of the Board.

Finally, I should like to recognize with thanks the devoted and efficient work of

several continuing members of the *Journal* staff, Lyndis Rankin and Margaret Yanchossek in Pittsburgh, and Ann G. Vollmer and Anita Makler at Stanford.

MOSES ABRAMOVITZ, *Managing Editor*

# Report of the Director

## *Job Openings for Economists*

For the second consecutive year, the number of new jobs listed declined from the previous year. Last year (1981), 1,754 new vacancies were advertised; this year only 1,659 new jobs were listed—a decline of slightly more than five percent from 1981. Both academic and nonacademic listings decreased although the percentage decline in nonacademic jobs was greater. Table 1 shows total listings (employers), total jobs, new list-

ings, and new jobs, by type (academic or nonacademic) for each issue of *JOE* in 1982.

Universities with graduate programs and four-year colleges continue to be the major sources of job listings. Together they constitute about 83 percent of total employers. Table 2 shows the number of employers by type for each 1982 issue. The largest percentage decline in employer type was State/Local Government. In 1981, fourteen

TABLE 1—JOB LISTINGS FOR 1982

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	83	169	69	137
April	65	106	58	93
June	26	39	23	36
August	33	68	31	62
October	121	293	106	278
November	109	269	109	269
December	184	620	123	275
Subtotal	621	1,564	519	1,150
Nonacademic				
February	10	49	16	35
April	16	85	14	75
June	15	73	11	53
August	16	69	10	64
October	23	103	19	96
November	14	61	14	61
December	30	160	23	125
Subtotal	124	600	107	509
TOTALS	745	2,164	626	1,659

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1982

Issue	Four-Year Colleges	Universities with Graduate Programs	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	34	49	2	1	3	—	2	2	93
April	34	31	4	1	5	2	2	2	81
June	15	11	4	—	4	1	3	3	41
August	11	22	2	1	4	2	5	2	59
October	38	83	6	—	7	2	6	2	144
November	43	65	4	1	2	—	8	—	123
December	58	126	8	1	8	1	9	3	214
TOTALS	234	387	30	5	31	8	35	14	745

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1982

Fields <sup>a</sup>	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	69	61	31	31	129	100	194	615
Growth and Development (100)	20	24	14	19	36	19	49	181
Econometrics and Statistics (200)	39	34	24	18	64	40	79	258
Monetary and Fiscal (300)	35	30	16	20	75	36	105	317
International Economics (400)	22	17	10	16	34	25	53	177
Business Administration, Finance, Marketing and Accounting (500)	27	21	15	8	44	23	66	204
Industrial Organization (600)	12	14	14	8	35	17	46	146
Agriculture and Natural Resources (700)	10	10	11	8	23	13	26	101
Labor (800)	10	13	7	12	27	23	41	133
Welfare and Urban (900)	16	10	8	11	26	15	46	132
Related Disciplines (A00)	4	2	—	1	9	8	12	36
Administrative Positions (B00)	10	4	2	3	13	6	10	48
TOTALS	274	240	152	155	515	325	727	2,388

<sup>a</sup>Fields of specialization codes are from the *Journal of Economic Literature*.

state or local government agencies listed jobs; this year, only five.

The field of specialization most in demand continues to be general economic theory. Generalists with a strong background in mathematics and statistics appear to be the type of economist that employers are seeking. The applied area of specialization seems to be of secondary importance. Table 3 shows the number of citations by field of specializa-

tion. General economic theory (000) led, followed by monetary and fiscal (300) and econometrics and statistics (200). This pattern has prevailed for the past several years.

Violet Kohsman is almost solely responsible for the publication and distribution of *JOE*. I wish to express my great gratitude for the excellent job she continues to do.

C. ELTON HINSHAW, *Director*

## International Economic Association

The Seventh IEA World Congress will be held in Madrid, Spain, September 5-9, 1983. The general topic will be "Structural Change, Economic Interdependence and World Development." The Executive Committee of the IEA has approved the program for the Congress, including the authors of the seven lectures invited for the plenary sessions, the first discussants for these papers, and the names of the organizers of the five panel discussions and the sixteen specialized sessions.

Conferences on the following topics were undertaken in 1982: "New Developments in the Theory of Market Structures" (chaired by Joseph Stiglitz), "Military Expenditures and Economic Growth and Fluctuations" (chaired by Christian Schmidt), "Structural Adjustments in Trade-Dependent Advanced

Countries" (chaired by Karl G. Jungenfelt), "Monetary Theory and Economic Institutions" (chaired by Marcello de Cecco and Jean-Paul Fitoussi), "Economics of Alternative Energy Sources" (chaired by Pierre Maillet and Shigeto Tsuru), and "Recent Developments in Economics, with Special Reference to International Economic Relations" (chaired by Victor Urquidi).

In addition to the Seventh World Congress, two conferences are currently planned for 1983. H. M. A. Onitiri is preparing "Structural Change, Economic Interdependence and African Development" to be held in Addis Ababa, Ethiopia in June, and Jesus Estanislao is organizing "Economic Interdependence: Perspectives from Developing Countries" for May 1983 in Manila, Philippines.

## Report of the Committee on Economic Education

After ten years of successful publication as an biannual, the *Journal of Economic Education* has expanded its coverage and size, and will now appear as a quarterly under the new editorship of Donald Paden, University of Illinois-Urbana. Associate editors are: William E. Becker, Karl E. Case, George Dawson, and Kalman Goldberg. The *Journal*, which has to date emphasized economic education research, will continue this focus, but will, in addition, include articles on pedagogy: what and how to teach, as well as materials and aids for classroom teachers. The first issue of the quarterly will appear in January 1983. Funding to support the transition to the expanded journal has been provided by the J.M. Foundation and the estate of Robert V. Horton.

The fourth of the series of Teacher Training Programs (TTP) was held at Harvard University in June under the directorship of Jeff Wolkowitz. The TTP is a packaged course, complete with *Resource Manual* and video tapes, designed to train faculty and graduate students in the techniques of teaching economics. Previous summer workshops

have been held at the Universities of Wisconsin, Indiana, and North Carolina-Chapel Hill. These highly successful workshops have been funded by the Lilly Endowment, Inc. The fifth workshop, planned for the summer of 1983 and funded by the Borg-Warner Foundation and Citibank, will emphasize a "regional training" format. Selected past TTP participants will be trained to sponsor and hold local workshops, thereby expanding the scope of the program and cutting its travel and related costs.

The study of the economic major, carried out by John Siegfried (Vanderbilt) and sponsored by the Alfred P. Sloan Foundation, is presently in its final stages of completion. The study has taken a census of economics departments to identify the nature of their major and course offerings, as well as obtaining data from graduating majors at selected colleges and universities. Several papers have been written and presented from this study. A public use tape of the data will be made available in 1983.

ALLEN C. KELLEY, *Chair*

## Report of the Representative to the National Bureau of Economic Research

Research at the National Bureau of Economic Research is organized in nine programs: Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel), Health Economics (Victor Fuchs and Michael Grossman), Law and Economics (William Landes), and Productivity and Technical Change (Zvi Griliches).

Work continued during 1982 on several large-scale projects, which bring together researchers from several of these programs. John Shoven is overall director of the Bureau's pensions project. A major phase of that project (directed by Zvi Bodie) dealing with the financial aspects of pensions concluded during 1982. The Bureau's work (directed by David Wise) on the effects of pensions on labor market and retirement decisions is continuing.

Completed during 1982 was the project under the direction of Mervyn King which compared capital taxation in Germany, Sweden, the United Kingdom, and the United States. Also nearing conclusion is that phase of Benjamin Friedman's research project on the changing roles of debt and equity finance which deals with corporate capital structures in the United States.

Work continues at the Bureau on simulation of the effects of changes in tax policy on tax revenue and on individual behavior. The Bureau has recently established a center for tax research to provide a more permanent institutional environment for NBER research in taxation. Zvi Griliches and M. Ishaq Nadiri are leading a study of research and development. The project on minority youth unemployment, led by Richard Freeman, continued in 1982.

Also in 1982, two major new research projects were begun, each encompassing a number of important research topics. One, centering on the role of the Government

Budget and the Private Economy, will consist of major efforts in the following areas: the impact of taxation on such behavior as labor supply and charitable contributions; measuring and analyzing the growth of government spending; an analysis of the impact of transfer programs, studies of public sector payrolls (directed by David Wise), and the impact of public sector unionization (directed by Richard Freeman); an analysis of state and local government budgets (directed by Harvey Rosen); and an analysis of government debt and deficits and their impact on the private sector (directed by David Bradford and Benjamin Friedman). The second project, Productivity and Industrial Change in the World Economy, likewise has several major parts. William Branson and David Richardson are directing a project on international economic policy. Research on trade relations and trade policy is directed by Anne O. Krueger and Robert Baldwin. Richard Marston has begun a project on international macroeconomic coordination. Colin Bradford is directing a study of trade relations with Asian countries. Jeff Sachs and Barry Eichengreen are studying domestic responses to changes in trade patterns. Jacob Frenkel is directing research on exchange rate changes.

Bureau conferences (and organizers) in the United States and abroad in 1982 included: "Exchange Rates" (Richard Marston and William Branson); "Classical Gold Standard" (Anna Schwartz); "Financial Aspects of the U.S. Pension System" (Zvi Bodie, John Shoven, and David Wise); "International Tax Comparison" (Mervyn King); "Inflation and Business Cycle Fluctuations" (Robert Barro and Robert King); "Transfer Payments" (Marilyn Moon); "Development of the American Economy" (Claudia Goldin); "International Seminar on Macroeconomics" (Robert Gordon and George de Menil); "Public Economics" (Robert King); "Fourth Annual Research Conference"; "Incentive Effects of Government Spending" (David

Bradford); "Trade Relations" (Anne Krueger and Robert Baldwin).

In 1982, the following NBER books were published by the University of Chicago Press: *The Youth Labor Market Problem: Its Nature, Causes, and Consequences* (Richard B. Freeman and David A. Wise, eds.); *Trade and Employment in Developing Countries, 2: Factor Supply and Substitution* (Anne O. Krueger, ed.); *Economic Aspects of Health* (Victor R. Fuchs, ed.); *The Economics of Information and Uncertainty* (John J. McCall, ed.); *Import Competition and Response* (Jagdish Bhagwati, ed.); *The Changing Roles of Debt and Equity in Financing U.S. Capital Formation* (Benjamin M. Friedman, ed.); *Monetary Trends in the United States and United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867-1975* (by Milton Friedman and Anna J. Schwartz); *The National Balance Sheet of the United States, 1953-80* (by Raymond W. Goldsmith); and *The U.S. National Income and Product Accounts: Selected Topics* (Murray F. Foss, ed.).

Over 180 participants, representing sixty-five universities and other organiza-

tions in the United States and abroad, met in Cambridge in July and August for the Bureau's fifth annual Summer Institute. Six NBER programs and projects held workshops and seminars: Economic Fluctuations, Financial Markets and Monetary Economics, International Studies, Labor Studies, Productivity, and Taxation.

The Business Cycle Dating Group met January 6, 1982 and determined that July 1981 was a peak in the business cycle. The group has not yet identified a trough to signify the end of the recession.

The National Bureau's President, Martin Feldstein, was appointed Chairman of the President Council of Economic Advisers in September 1982. He was succeeded as Bureau President by Eli Shapiro. Further information on Bureau activities is available in the *NBER Reporter*, from David G. Hartman, Executive Director, NBER, 1050 Massachusetts Avenue, Cambridge, Massachusetts 02138, or from the undersigned at Johns Hopkins University.

CARL F. CHRIST, *Representative*

## Report of the Representative to the U.S. National Commission for UNESCO

The main activity of the U.S. National Commission for UNESCO in 1982 was a special meeting on "A Critical Assessment of U.S. Participation in UNESCO." This meeting, held on June 1-3 at the University of South Carolina, was convened at the instance of the Commission's new chairman, James Holderman, with the approval of the Commission's Executive Committee. It reflected at least in part a sense of tension and concern about the American relationship with UNESCO. This was the first comprehensive assessment of U.S. participation in UNESCO since 1960.

Ninety-three people, representing academic, business, and government communities participated in the meeting. A background paper, "The United States and UNESCO: Is the Past Prologue?," prepared by Lawrence Finkelstein, a veteran member of the Commission and student of UNESCO, pointed out that there have been difficulties in the relationship between the United States and UNESCO over a long period, but they have intensified in recent years because of the increase in UNESCO's role regarding issues that affect U.S. domestic interests that have political influence (for example, issues of communication and information and of Arab-Israeli relations in the Middle East), and by loss of the dominance that the United States once enjoyed. The paper provided a balanced assessment of UNESCO, the arguments for and against U.S. participation, and the alternative courses open to the United States. After consideration of this paper by the members individually and in a plenary session, discussions in five working groups, and a final plenary session at which reports of all the working groups were presented, the Commission agreed on several conclusions.

First it was recommended by all the working groups that the United States not only continue its membership in UNESCO, but increase the effectiveness of its participation in UNESCO's work. The main grounds for this recommendation were that the U.S. ben-

efits politically and psychologically by participation, and also, although this was regarded as more incidental, economically through expenditure of funds in the United States. Several groups concluded that the United States had not been sufficiently active in initiating program proposals. Another theme running through the group reports was the need for the Department of State to interact more effectively with private organizations, especially those most concerned with UNESCO and its interests, and for the National Commission to play a more active role as intermediary between the U.S. government and the professional constituencies represented by the members.

The Social Sciences Committee of the Commission considered the medium-term plan covering the years 1984-89 proposed by the Secretariat of UNESCO, which was to be discussed at an extraordinary session of the Organization's General Conference late in 1982. The Committee supported the Secretariat's view that UNESCO's work in the natural sciences and human sciences should be brought together in the same context. Secondly, it stressed the need for UNESCO to give priority to building mechanisms to feed back the findings and techniques of social science research taking place in developing countries into the discipline as a whole. (There were traces of this idea in the Secretariat's program but too little provision, the Committee thought, for translating the idea into practice.) Third, the Committee expressed disappointment at the absence of any provision for collaborative research in the social sciences between developing and developed countries, since it is aware that there are highly competent social science researchers in many regions of the world and there is an interest in such collaborative undertakings. Finally, the Committee questioned the Secretariat's proposal that UNESCO give greater emphasis to the management sciences, since the ILO already has a program in this field and the funds could be

better used by UNESCO in other disciplines. The Committee stressed the essentiality of the traditional disciplines of political science, sociology, and economics as more universal than some others that the Secretariat proposed to emphasize more and because they bear most directly on the process of development. The U.S. Ambassador to UNESCO, Jean Gerard, in her statement on the Medium-Term Plan before the Executive Board of UNESCO in September 1982, included the first three of these points in her major policy statement.

Apart from these activities related to the social sciences, the National Commission sponsored a meeting of U.S. experts in various aspects of cultural policy. This meeting was called to consider issues and strategies likely to arise at the Second World Conference of UNESCO on Cultural Policy, held in Mexico City, July 25–August 6, 1982, in anticipation of this conference being strongly political and involving attacks on U.S. “cultural imperialism.”

Experts in the field of education held a meeting in Vienna, March 31–April 2, 1982, to discuss the possibility of joint studies in education. Two members of the U.S. National Commission participated in this meeting. The result of the meeting was agreement to conduct fourteen joint studies. The United States will have a coordinating role jointly with Canada in a study of what was called “New Technologies in Information and Communications and Their Impacts on Education,” and will also have a coordinating role in a study of “Higher Education Institutions as Centers of Lifelong Learning”; both are three-year studies. It is planned to have the working groups meet four times during this period. Any member of the AEA wishing to contribute to either of these studies should communicate with me.

Some major projects under international programs to develop communication were also initiated as a result of this meeting.

WALTER S. SALANT, *Representative*

## Report of the Committee on U.S.–China Exchanges

Much has happened in exchanges in economics between the United States and the People's Republic of China since my Report a year ago (see this *Review*, May 1982, p. 429). The news is essentially good.

Some bad news first. Under the sponsorship of the Committee on Scholarly Communication with the People's Republic of China, a team of U.S. economists, to be headed by Randy Barker and Robert Dernberger, was to visit China to study the recent changes in agricultural economic policy in China. The host of this team's visit was to be the Chinese Academy of Social Sciences. Because the Chinese Academy could not agree to some of the requests from the team concerning their itinerary, the members of the team, after serious deliberations, decided to postpone the visit, leading to its eventual cancellation. In the meantime, however, other groups of economists have visited China, including a group from the National Bureau of Economic Research in May/June 1982 and a group from The Brookings Institution in June 1982.

Second, a volume entitled *Essays on the Economies of China and other Developing Countries by Foreign Economists* (in Chinese) was published in early 1982 by the Editorial Board of *Economic Research*, a journal of the Institute of Economics of the Chinese Academy of Social Sciences. This volume consists of essays by American economists, including Irma Adelman, Kenneth Arrow, Gregory Chow, Robert Dorfman, Ronald Duncan and Helen Hughes, Dwight Perkins, Joseph Stiglitz, Paul Streeten, and Laura D'Andrea Tyson.

Third, and related to the second, is the proliferation of economics journals in China, published mostly by Chinese universities. A glance through these journals reveals that more and more articles deal with the tools of Western economic analysis. From these journals, a university student in China can get a fair exposure to Western economic concepts, though not a systematic treatment.

Fourth, from my own visit to China in June/July 1982 for the purpose of lecturing and exchanging ideas with economists in five major universities located from Canton to Peking, I have found that many Chinese universities are expanding their teaching and research in economics. Major economics departments have introduced courses in micro- and macroeconomic theory and econometrics, while some are planning to have Western economics and econometrics as fields of concentration. Chinese scholars are learning these subjects rapidly and the rate of growth is impressive. A Society of Quantitative Economics and Econometrics was founded at a meeting of some 200 economists which took place in Xian in March 1982. Other organized activities, some affiliated with associations in industrial engineering and systems science, are being contemplated to promote research and communication in quantitative economics.

Fifth, during the past year many more graduate students and visiting scholars in economics have come to the United States from the People's Republic of China than in past years, while American economists have continued to go to China to lecture and to visit Chinese economics institutions. Visits by American economists and scholars in general have been facilitated by the increase in housing for foreign visiting scholars in China. For example, Peking University completed a residential hall with some 250 rooms to accommodate foreign visitors. A distinguished American economist travelling to the Far East can arrange a side visit to China fairly conveniently through a Chinese university.

In summary, economics exchanges between the United States and the People's Republic of China have expanded and will continue to do so, mainly through the decentralized initiatives of economists and academic institutions in both countries wishing to promote them.

GREGORY C. CHOW, *Chair*

## Policy and Advisory Board for the Economics Institute

The Board held its formal meeting December 27, 1982 in New York in connection with the ASSA meetings. Several Board members also visited the Institute during the summer of 1982.

Nineteen eighty-two was another successful year for the Institute. Total enrollment was 552, up from 546 in 1981, and average length of stay was down slightly, from nineteen weeks in 1981 to seventeen weeks in 1982. The shorter stays occurred during the academic year may result from the world-wide recession.

The twenty-fifth consecutive summer session of the Institute was held in 1982. Summer enrollment was 36 in 1958, the first year of the Institute; and 418 in 1982. During its first quarter-century, the Institute's program has not only grown rapidly in size, but also greatly improved its quality. Conversion to a year-round program in 1976, was a watershed in both respects. It permitted recruitment of year-round faculty for both subject matter and English instruction. The result has been much more systematic

and innovative instruction which integrates teaching of language and subject matter.

Many countries that provide Institute students are suffering from the recession and foreign exchange shortages that are affecting much of the developing world. In addition many U.S. universities which Institute alumni attend are experiencing financial problems which may make it more difficult for them to finance study at the Institute. We expect 1983 to be a year of some financial stress for the Institute.

During 1982, Axel Leijonhufvud, University of California-Los Angeles, completed his term on the Board; John Moroney, Texas A&M University, joined the Board.

A variety of activities are being planned to celebrate the Institute's twenty-fifth birthday. At the ASSA meeting, a group of his friends held a luncheon to honor Wyn Owen for his twenty-five years as Director of the Institute.

WYN F. OWEN, *Director*

EDWIN S. MILLS, *Chairman*

## The Committee on the Status of Women in The Economics Profession

The first decade of the Committee on the Status of Women in the Economics Profession (CSWEP) has seen little, if any, progress for women economists in academe. According to a matched sample of forty Ph.D. granting departments, the number of women full professors increased from nine women in 1978-79 to ten women in 1981-82. This net gain of one woman over the four-year period did not represent a percentage gain. Instead, the percentage held constant with women comprising 1.8 percent of all persons holding full professor rank. The percentage of women associate professors did, however, increase from 4.2 percent in 1978-79 to 5.4 percent in 1981-82 and represented a net increase of six women in these departments. Counteracting somewhat the gains at the associate professor level is a slight decline in the percentage of women assistant professors from 13.1 percent to 12.6 percent in this four-year sample. This decline occurred despite a near doubling of both the number and percentage of women granted Ph.D.s in economics over this period. Even the small net gain represented at the associate professor level may not reflect accurately our status at all departments, since there may be some self-selection among the responding schools. The matched sample may well reflect schools that have a better record with respect to women than the nonanswering departments.

Viewing status from the vantage point of the top academic departments of economics, only M.I.T. among the top six departments (as ranked by F. M. Boddy in December 1981) has a woman economist at the tenured level. The economics departments at Chicago, Harvard, Stanford, Princeton, and Yale have no tenured women, although one of these schools has a tenured woman economist in a noneconomics department. Indeed, in 1981-82, these departments had fewer women in the assistant professor rank than in any year in the recent past. Two of the departments have no women economists in any professorial rank. Thus, women are repre-

sented more poorly in the top economics departments than they were four years ago.

In other dimensions, each year brings with it one or two more visible gains for women economists. Women economists have been appointed to a number of senior positions in government, including one at the cabinet level. A few women economists have moved into officer ranks in business, banking, research and consulting firms or into administrative positions in academe. Women economists have become more widely represented in the annual meetings. In the 1982 annual meetings, for example, over two dozen sessions were chaired by women, and roughly 45 percent of the sessions had a woman participating. In addition, a number of women have been elected to positions of responsibility in the American Economic Association. Thus, in honorific dimensions within academe and in the nonacademic job market, women economists are making some progress. In the bread-and-butter dimensions of jobs in academe, both in being hired and promoted, women do not appear to be making progress and the base remains at a low level.

As we embark on the second decade of CSWEP's existence, CSWEP must make renewed efforts to fulfill its desired function of improving the status of women in the academic work environment. To respond to this challenge, CSWEP has a new chair, Barbara Bergmann. Barbara has a history of advocacy for women which should serve her well as she faces the task ahead. The statistics indicate that economics is attracting larger numbers of women Ph.D.s. In the four-year matched sample of forty Ph.D.-granting departments, the number of women receiving their doctorates almost doubled (from 22 to 42). These figures highlight the fact that the female faculty which now exists is relatively junior. The statistics also show, however, that the percentage of women Ph.D.s who turn toward academe is below that for men and has declined from 58.8

percent in 1978-79 to 52.0 percent in 1981-82. This trend may be due to a recognition by women that opportunities for promotion are poor in academe. Certainly, there must be a strong thrust in CSWEP's second decade to insure that academic women have the same opportunities for career advancement as do their male counterparts.

As my term as chair comes to a close, I feel the greatest accomplishment for the Committee during my period of stewardship has been the publication of the CSWEP Roster. Nancy Ruggles has overseen the production and distribution of the Roster of Women Economists. The format is attractive and easy to use. The Roster is useful in locating women by field of research interest as well as by employer. For example, it reveals that over a dozen women economists are employed by the Board of Governors of the Federal Reserve System and by the World Bank. The Roster lists phone numbers and makes it easy for members to contact other women when giving a seminar at a distant university. The Roster thus serves a networking function in addition to its value in searching for women for job opportunities or for appearances on programs. I am glad that this new service from CSWEP has been added to our thrice-yearly newsletter and to our sponsorship of sessions on women's issues at all of the major regional and national association meetings.

I wish to thank all of the members of my Committee for their support during these last three and a half years. Without their willingness to share responsibility, we could not have accomplished as much as we have. I am particularly grateful to Nancy Ruggles for her willingness to serve a second term as we seek to find ways in which care of the Roster can be less burdensome for the Committee. I am grateful as well to Louise Curley for the fine job she did with the newsletter, and wish Aleta Styers well as she assumes this task for the next three years. Jean Shackelford departs the Committee after having initiated workshops on specific topics related to econometric methods and economic theory, and leaves us with funds for one more such workshop, which her replacement Cordelia Reimers should appreciate. Bob Eisner leaves

the Committee this year, after loyal service in which he provided much sensible advice as well as liaison services with the Econometrics Society. Joseph Pechman has agreed to replace him, and is already at work in seeking funding for a joint Brookings-CSWEP conference on men and women in the work place. The remaining members of the committee should serve to provide continuity for Barbara Bergmann during the transition period as she assumes her new position.

### I. CSWEP Activities

CSWEP has become involved in two new initiatives this year. One of these is the planning for a joint conference with The Brookings Institution. Joe Pechman and Claire V. Brown are working together on this project. The idea would be to provide a forum for thoughtful and original research on issues relating to men and women in the work force, with the papers then coming out in a Conference Volume. CSWEP is quite excited about this new venture. We hope that the topics will cover a broad range of research on work place issues, and that young researchers and new research approaches will have an opportunity to receive constructive criticism from senior scholars. CSWEP was disappointed that the National Bureau of Economics Research backed away from participation in such a project, but is quite delighted at the interest shown by Brookings. We are looking forward to a warm and productive partnership.

A second initiative is one suggested by Simmons College. In this plan, there would be a collaboration between a leading women's college focusing on careers for women and the American Economic Association, dedicated to the improvement of information about the economics profession nationally. The project would use library services to develop materials and information packets about the economics profession, and would promote economic career forums throughout the nation for women undergraduates.

Other CSWEP activities this year have involved consolidation and continuation of ongoing projects. We continue an active collaboration, through Gail Wilensky, with

Washington Women Economists and with the Federation of Organizations of Professional Women. We continue to sponsor sessions at annual meetings of the American Economic Association (AEA) and of the regional associations. At the annual meetings in New York in December 1982, CSWEP sponsored two sessions—one on "Women and Health" and the other on "Comparable Worth: Does it Have any Economic Meaning." In addition, our annual business meeting continues to attract a large number of women, whose suggestions are greatly appreciated.

As in previous years, CSWEP continues to provide a flow of information to women economists. Our thrice-annual newsletter, under the direction of Louise Curley, presents calls for papers, summarizes committee activities, offers a plethora of announcements, publications, and generally useful in-

formation for women in our profession. The chief new addition to the newsletter this year is a section reporting on recent meetings and conferences of interest to women. Our thought here has been to permit all women to learn of the coalition strategies and concerns of a broad base of professional women, including those from other academic disciplines as well as those in economics.

This year has marked the publication of our second Roster of Women in Economics. The latest Roster came out in September 1982, timed to be available from the beginning of the recruiting season. As was true last year, entries are in plain English, not codes, and include name, address, and telephone numbers, publications, fields of specialization, and current research interests. Indexes by speciality and location also appear, along with a new index on place of employment. The Roster Directory has been distrib-

TABLE 1—DISTRIBUTION OF FULL-TIME FACULTY, BY TYPE OF INSTITUTION, ACADEMIC YEAR, 1981-82

	Chairman's Group			Other Ph.D.			Only M.A. Departments			Only B.A. Departments		
	Female			Female			Female			Female		
	Total	No.	Percent	Total	No.	Percent	Total	No.	Percent	Total	No.	Percent
Existing												
Professor	643	10	1.6	445	10	2.2	167	6	3.6	204	22	7.7
Associate	252	17	6.7	244	13	5.3	161	15	9.3	307	19	6.2
Assistant	353	44	12.5	223	33	15.8	148	20	13.5	395	57	14.4
Instructor	40	5	12.5	24	5	20.8	31	8	25.8	135	28	20.7
Other	32	6	18.7	21	6	28.6	5	1	20.	39	6	15.4
New Hires												
Professor	7	0	—	8	0	—	5	0	—	15	0	—
Associate	9	1	1.1	15	2	13.3	4	1	25.	20	4	20.
Assistant	76	6	7.9	57	6	10.5	32	10	31.2	148	15	10.1
Instructor	11	3	27.3	4	1	25.	12	3	25.	60	21	35.
Other	11	1	9.1	5	3	60.	1	0	—	16	11	68.7
Promoted to Rank (1980-81)												
Professor	20	1	5.	24	2	8.3	13	1	7.7	17	2	11.8
Associate	41	1	2.4	21	2	9.5	8	3	37.5	20	1	5.
Assistant	10	1	10.	0	0	—	1	0	—	12	2	16.7
Tenured at Rank (1980-81)												
Professor	2	0	—	6	0	—	2	0	—	2	0	—
Associate	25	1	4.	20	1	5.	7	0	—	14	0	—
Assistant	0	0	—	1	0	—	1	0	—	7	1	14.3
Not Rehired												
Professor	22	0	—	11	0	—	3	0	—	8	—	12.5
Associate	5	0	—	6	0	—	9	2	22.2	11	0	—
Assistant	45	6	13.3	16	4	25.	14	1	7.1	38	7	18.4
Instructor	9	2	22.2	6	0	—	6	1	16.7	18	4	22.2
Other	1	0	—	3	0	—	0	0	—	9	2	22.2

uted to all major university departments and is sent to all dues-paying members and associate members as a benefit of membership.

## II. Status of Women Economists in Academe

Each year, the Universal Academic Questionnaire is distributed by the AEA to all department chairman and the responses are tabulated by Charles Scott of Marquette University. It is the most comprehensive source of information on the academic labor market in economics. However, responses are voluntary and in recent years, its information is often provided by only two-thirds or less of academic departments. Annual comparisons are difficult because the responding institutions vary from year to year. Hence, the data provided in Tables 1-6 provide a useful but not fully accurate view of the role of women in the academic labor market when

compared with their counterpart tables published in previous CSWEP reports. Fortunately, we do have this year a matched sample of forty Ph.D.-granting departments, who have for each of the last four years all consistently responded to the questionnaire. The results of this sample are given in Table 7, and provide a consistent comparative snapshot of our status in 1978-79 and in 1981-82. Whether or not this snapshot is fully accurate depends, unfortunately, on whether there is some self-selection in the responses. It may be that the departments with the poorest records with respect to women tend not to report.

Table 1 provides a summary of the distribution of academic jobs at the beginning of the academic year 1981-82. It presents information for four types of departments: the Chairman's Group; other Ph.D. departments; M.A. departments; and B.A. depart-

TABLE 2—PREVIOUS ACTIVITY OF NEW HIRES AND CURRENT ACTIVITY OF THOSE NOT REHIRED  
BY TYPE OF INSTITUTION AND SEX, ACADEMIC YEAR, 1981-82

	Previous Activity of New Hires				Current Activity of Not Rehired			
	Male		Female		Male		Female	
	No.	Percent	No.	Percent	No.	Percent	No.	Percent
Chairman's Group	102	100.0	13	100.0	67	100.0	17	100.0
Faculty	26	25.5	3	23.0	45	67.1	9	52.9
Student	60	58.8	7	53.8	1	1.5	4	23.5
Government	1	1.	1	7.7	3	4.5	1	5.9
Bus., Banking, Research	2	2.	1	7.7	9	13.4	2	11.8
Other	13	12.7	1	7.7	9	13.4	1	5.9
Other Ph.D.	82	100.0	13	100.0	33	100.0	4	100.0
Faculty	23	28.4	3	23.0	18	54.5	2	50.
Student	39	47.6	5	38.5	1	3.0	0	—
Government	4	4.9	0	—	2	6.1	0	—
Bus., Banking, Research	9	11.	2	15.4	5	15.2	2	50
Other	7	8.5	3	23.7	7	21.2	0	—
M.A. Departments	45	100.0	15	100.0	21	100.0	6	100.0
Faculty	19	42.2	6	40.	9	42.9	1	16.7
Student	14	31.1	7	46.7	2	9.5	0	—
Government	1	2.2	0	—	2	9.5	1	16.7
Bus., Banking, Research	9	20.	2	13.3	4	19.0	4	66.7
Other	2	4.4	0	—	4	19.0	0	—
B.A. Departments	143	100.0	44	100.0	79	100.0	43	100.0
Faculty	46	32.3	8	18.2	44	55.7	8	18.6
Student	70	49.	24	54.4	8	18.2	3	7.
Government	4	2.8	4	9.1	2	2.5	16.	37.2
Bus., Banking, Research	16	11.2	3	6.8	16	20.3	14	32.6
Other	7	4.9	5	11.4	9	11.4	2	4.7

TABLE 3—DISTRIBUTION OF SALARY FOR WOMEN FACULTY BY TYPE OF DEPARTMENT AND TIME IN RANK, ACADEMIC YEAR, 1981-82

Relative Salary for Rank	All Women		Time in Rank			
	Number	Percent	Total Percent	Above Median	At Median	Below Median
All Departments	307	100.0				
Salary above Median	104	33.9	100.0	43.2	34.6	22.1
Salary at Median	109	35.5	100.0	11.9	72.4	15.6
Salary below Median	94	30.6	100.0	15.9	19.1	64.8
Ph.D., Chairman's	85	100.0				
Salary above Median	26	30.6	100.0	34.6	50.	15.3
Salary at Median	28	32.9	100.0	10.7	82.1	7.1
Salary below Median	31	36.5	100.0	19.3	29.0	41.6
Ph.D., Other	64	100.0				
Salary above Median	25	39.0	100.0	56	28	16
Salary at Median	24	37.5	100.0	25	54.1	20.8
Salary below Median	15	23.4	100.0	13.3	26.6	60
M.A. Departments	55	100.0				
Salary above Median	17	30.9	100.0	41.1	35.2	23.5
Salary at Median	16	29.0	100.0	12.5	56.2	31.2
Salary below Median	22	40.0	100.0	22.7	4.5	72.7
B.A. Departments	103	100.0				
Salary above Median	36	35.	100.0	41.6	27.7	30.5
Salary at Median	41	39.8	100.0	4.8	82.9	12.2
Salary below Median	26	25.2	100.0	7.6	15.3	76.9

ments. The Chairman's Group consists of sixty-five departments that focus on research and the training of Ph.D.s in economics. In terms of stature, it is generally agreed that academic appointments at a department within the Chairman's Group carry the most prestige. Thus, this discussion will tend to focus upon the role of women in the Chair-

man's Group as a bellwether for the entire economics profession. The other Ph.D. granting departments focus primarily on undergraduate education, but also have a viable Ph.D. program. The M.A. departments, similarly, have their primary focus upon undergraduate education, but also have a Master's program. Finally, the B.A. departments are

TABLE 4—DEGREES GRANTED IN ECONOMICS BY TYPE OF DEPARTMENT AND SEX, ACADEMIC YEAR 1981-82

Number of:	All Depts.	Ph.D. Departments			M.A. Depts.	B.A. Depts.
		Total	Chairman's	Other		
Departments	325	83	44	39	43	199
Ph.D.s	795	795	608	187	-	-
Female	122	122	97	25	-	-
Percent Female	15.3	15.3	16.0	13.4	-	-
M.A.s	1,359	1,113	632	481	246	-
Female	290	243	113	130	47	-
Percent Female	21.3	21.8	17.9	27.0	19.1	-
B.A.s	12,041	6,593	4,256	2,337	843	4,605
Female	3,967	1,933	1,223	710	226	1,808
Percent Female	32.9	29.3	28.7	30.4	26.8	39.3
Other	68	68	45	23	-	-
Female	17	17	7	10	-	-
Percent Female	25.0	25.0	15.5	43.5	-	-

exclusively concerned with undergraduate teaching.

According to Table 1, at the forty-four departments in the Chairman's Group who responded this year, one woman was brought in as a new hire at the associate level, and one woman was promoted to each of the ranks of associate and full professor, respectively. The gains were twice as good in the other Ph.D. departments, with two women appearing in each of the three categories. Thus, there does appear to be some gain for women in 1981-82. The overall percentage of women full professors remains at about two percent in the Ph.D. departments, which is about the level it has been throughout the last decade. The percentage of women associate professors appears to be improving somewhat.

Table 2 supplies some information about the previous activity of those who are newly hired and the present activity of those who have not been rehired. As in previous periods, there is some indication that women are more likely than men to choose nonacademic careers when they have not been rehired by their academic department.

Table 3 describes the salary distribution for women faculty by type of departments and time in rank. Table 3 indicates that women are doing worse in relative salary treatment in the Chairman's Group than in the other Ph.D. granting universities. For example, 30 percent of the women in the Chairman's Group received salaries above the median in contrast to 39 percent of the women in the other Ph.D.-granting universities. Of the women who do receive a salary above the median in the Chairman's Group, however, two-thirds have time in rank at or below the median so the rising stars, as it were, are receiving relatively favorable salary treatment at the Chairman's Group.

Table 4 displays the percentages of women obtaining degrees in economics. There continues to be a strong increase in the percentage of women majoring in economics at all degree levels. At the Ph.D. level, the percentage has almost doubled from 8 to 15 percent in the past four years.

Table 5 contrasts the occupational choices of men and women Ph.D.s in 1981-82. It reveals that roughly half of both men and women Ph.D.s entered the academic labor

TABLE 5—DISTRIBUTION OF ACTIVITIES OF NEW PH.D. DEGREES BY SEX AND TYPE OF DEPARTMENT, ACADEMIC YEAR 1981-82

	All Depts.		Chairman's Group		Other Ph.D. Depts.	
	No.	Percent	No.	Percent	No.	Percent
All Ph.D.s	561	100.0	411	100.0	150	100.0
Education	289	51.5	209	50.9	80	53.3
Government	45	8.0	37	9.0	8	5.3
Bus., Banking, Research	67	11.9	49	11.9	18	12.
Int'l. Emp. Outside U.S.	129	23.	100	24.3	29	19.3
Other	31	5.5	16	3.9	15	10.
Male Ph.D.s	504	100.0	368	100.0	136	100.0
Education	261	51.8	187	50.8	74	54.4
Government	40	7.9	33	9.	7	5.1
Bus., Banking, Research	53	10.5	40	10.9	13	9.6
Int'l. Emp. Outside U.S.	125	24.8	97	26.4	28	20.6
Other	25	5.	11	3.	14	10.3
Female Ph.D.s	57	100.0	43	100.0	14	100.0
Education	28	49.1	22	51.1	6	42.9
Government	5	8.8	4	9.3	1	7.1
Bus., Banking, Research	14	24.6	9	20.9	5	35.7
Int'l. Emp. Outside U.S.	4	7.0	3	7.	1	7.1
Other	6	10.5	5	11.6	1	7.1

TABLE 6—DISTRIBUTION OF PH.D. STUDENT SUPPORT, BY TYPE OF SUPPORT, SEX, AND DEPARTMENT  
ACADEMIC YEAR 1981-82

	All Ph.D. Depts.		Chairman's Group		Other Ph.D. Depts.	
	No.	Percent	No.	Percent	No.	Percent
All Students	4,233	100.0	3,079	100.0	1,154	100.0
Tuition Only	220	5.2	177	5.7	43	3.7
Stipend Only	404	9.5	182	5.9	222	19.2
Tuition + Stipend	1,755	41.5	1,377	44.7	378	32.8
No Support	1,075	25.4	807	26.2	268	23.2
No Record	779	18.4	536	17.4	243	21.0
Male Students	3,395	100.0	2,514	100.0	881	100.0
Tuition Only	179	5.3	145	5.8	34	3.9
Stipend Only	325	9.6	140	5.6	185	21.
Tuition + Stipend	1,419	41.8	1,122	44.6	297	33.7
No Support	869	25.6	642	25.5	227	25.8
No Record	603	17.8	465	18.5	138	15.7
Female Students	838	100.0	565	100.0	273	100.0
Tuition Only	41	4.9	32	5.7	9	3.3
Stipend Only	79	9.4	42	7.4	37	13.6
Tuition + Stipend	336	40.0	255	46.9	81	29.7
No Support	206	24.6	165	29.2	41	15.0
No Record	176	21.0	71	12.6	105	38.5

TABLE 7—MATCHED SAMPLE OF FORTY PH.D.-GRANTING DEPARTMENTS, 1978-82

	1978-79			1981-82		
	Total	Women	Percent	Total	Women	Percent
Distribution of Full-Time Faculty						
Professor	494	9	1.8	541	10	1.8
Associate	214	9	4.2	276	15	5.4
Assistant	289	38	13.1	317	40	12.6
Instructor	57	6	10.5	38	3	10.7
Other	51	8	15.7	48	8	16.7
Degrees Granted in Economics						
Ph.D.s	291	22	7.6	347	42	12.1
M.A.s	497	85	17.1	551	130	23.6
B.A.s	2,558	576	22.5	3,455	1,032	29.9

	1978-79				1981-82			
	Men	Percent	Women	Percent	Men	Percent	Women	Percent
Distribution of Activities of New Ph.D. Degrees								
Education	195	67.7	10	58.8	120	56.6	13	52.
Government	35	8.7	2	11.8	26	12.3	4	16.
Bus., Banking, Research	22	7.6	3	17.6	23	10.8	6	24.
Int'l. Exp. Outside U.S.	37	12.8	2	11.8	40	18.9	0	0.
Other	9	3.1	0	0.	3	1.4	2	8.
Distribution of Ph.D. Student Support								
Tuition Only	97	6.2	17	5.5	82	4.8	17	4.7
Stipend Only	288	18.5	50	16.1	160	9.3	44	12.2
Tuition + Stipend	680	43.7	126	40.5	793	46.0	179	49.7
No Support	424	27.3	94	30.2	356	20.6	63	17.5
No Record	66	4.2	24	7.7	334	19.3	57	15.8

market in 1981-82. The percentages of men and women going into government declined slightly from previous years, but remains in the usual 8-10 percent range. Nearly 25 percent of women Ph.D.s entered business, banking, and research firms, up substantially from previous years.

Table 6 continues to show that women are doing as well as men in graduate student support. No discernable evidence of decreased support for graduate study in general is yet in evidence.

Perhaps the most interesting evidence on the status of women appears in Table 7, which summarizes the results of a matched sample of forty Ph.D.-granting departments over the four academic years 1978-79 to 1981-82. These are all departments who consistently answered the questionnaire. The data indicates some growth in the number of women associate professors, as some of the bulge of women assistant professors hired in the mid-1970's receive promotions. The data indicate some decline in the percentage of

women in the assistant professor rank, despite the substantial increase in the percentage of women majoring in economics at all degree levels. The data suggest some movement away from academic employment by both men and women, and a consistently lower proportion of women than men in choosing academic careers. This would seem to indicate rational behavior on the part of women since expectation of success in the academic environment is not very high. In the area of graduate student support, the data indicate a greater level of support for all students in 1981-82 as compared with 1978-79, with women graduate students faring particularly well.

I must conclude then that the status of women economists in academe is plodding along at a whimper. A whimper is better than a silent standstill, but where, oh where, is our bang?

ELIZABETH E. BAILEY, *Chair*

## **Choice, Welfare, and Measurement**

*Amartya K. Sen*

"Sen's mastery in the fields of social choice, the foundations of welfare economics, and, more broadly, distributive ethics and the measurement problems associated with these fields is unquestioned. The selection of [twenty] articles fully reflects his work in this area . . . a number of papers are already classics."

—Kenneth Arrow

1983 440 pp. \$37.50

## **Banking on the Poor**

The World Bank and World Poverty

*Robert L. Ayres*

"An excellent history of the evolution of the World Bank under McNamara. . . . It provides an independent contribution to the debate on the causes of poverty and methods of alleviating it through economic growth . . . a superb work."

—Theodore H. Moran, Georgetown School of Foreign Service

1983 304 pp. \$17.50

## **Third World Multinationals**

The Rise of Foreign Investment from Developing Countries

*Louis T. Wells, Jr.*

1983 224 pp. \$25.00

## **Technology Choice in Developing Countries**

The Textile and Pulp and Paper Industries

*Michel A. Amsalem*

May 224 pp. \$27.50

## **Lectures on International Trade**

*Jagdish N. Bhagwati and  
T. N. Srinivasan*

This textbook by two eminent theorists of international trade presents the most integrated and ambitious treatment of the subject available to date.

July 464 pp. \$24.95

---

*Original in paperback*

## **Economic Interdependence and Flexible Exchange Rates**

*edited by Jagdeep S.*

*Bhandari and Bluford H.*

*Putnam, with Jay H. Levin*

1983 560 pp. \$15.00

(Cloth \$30.00)

*Write for our economics catalog*

28 Carleton Street  
Cambridge, MA 02142

# **THE MIT PRESS**

*The Regulation of  
Economic Activity Series*

---

**Folded, Spindled,  
and Mutilated**

Economic Analysis and *U.S. v IBM*

*Franklin M. Fisher,  
John J. McGowan, and  
Joel E. Greenwood*

*Foreword by Carl Kaysen*

This analysis of a major anti-trust suit uncovers the major flaws in the government's economic performance in attempting to define the nature of competition and monopoly.

June 440 pp. \$25.00

**Incentives for  
Environmental  
Protection**

*edited by*

*Thomas C. Schelling*

"The three thoughtful and well-informed case studies in this volume . . . provide well-documented explorations of how economic incentives compare with command regulations in dealing with specific and genuine environmental problems."—Robert Dorfman, Harvard University

May 384 pp. \$32.50

**United States Oil  
Pipeline Markets**

Structure, Pricing, and  
Public Policy

*John A. Hansen*

June 176 pp. \$22.50

---

*Now in paperback*

**Studies in Public  
Regulation**

*edited by Gary Fromm*

May 368 pp. \$15.00

**The Economics  
and Politics of  
Oil Price Regulation**

Federal Policy in the  
Post-Embargo Era

*Joseph P. Kalt*

May 338 pp. \$12.50

*Write for our economics catalog*

28 Carleton Street  
Cambridge, MA 02142

**THE MIT PRESS**

## **The Economics of Industrial Innovation**

Second Edition

*Christopher Freeman*

An expanded and updated account of the major economic problems associated with introducing new products and processes in manufacturing industries.

1983 258 pp. \$25.00

## **Worker Capitalism**

The New Industrial Relations

*Keith Bradley and*

*Alan Gelb*

May 208 pp. \$20.00

### *Organization Studies Series*

## **Control in the Police Organization**

*edited by Maurice Punch*

1983 368 pp. \$30.00

## **Disorganized Crime**

The Economics of the Visible Hand

*Peter Reuter*

1983 256 pp. \$17.50

## **An Introduction to Risk and Return from Common Stocks**

Second Edition

*Richard A. Brealey*

This considerably expanded edition of Brealey's popular work presents a brief nontechnical description of modern academic research on investment management, covering market efficiency, valuation, and modern portfolio theory.

May 192 pp. \$14.95

## **Money and Inflation**

*Frank Hahn*

"... a salutary reminder of the problem posed by the mere existence of money for the satisfactory construction of economic models. . . . More critically, however, Hahn picks away at the foundations of some recent economic analysis."—*London Review of Books*

*Review of Books*

1983 136 pp. \$12.50

---

*Now in paperback*

## **Studies in Business-Cycle Theory**

*Robert E. Lucas, Jr.*

April 316 pp. \$9.95

*Write for our economics catalog*

28 Carleton Street  
Cambridge, MA 02142

# **THE MIT PRESS**

## **A History of Economic Reasoning**

*Karl Pribram*

The final work of a noted economic scholar and policy maker is a unique methodological analysis of the development of economic thought. Karl Pribram's work spans the centuries from the late Middle Ages to the late 1950s, juxtaposing the often contradictory currents of economic theory with the evolution of Western thought over time. *A History of Economic Reasoning* not only clarifies the origins of the radically diverse teachings that compose the field of "economic science" but also bridges the gap between economics and the other social sciences. **\$42.50**

## **Environment, Natural Systems, and Development**

AN ECONOMIC VALUATION GUIDE

*Maynard M. Hufschmidt, David E. James, Anton D. Meister,  
Blair T. Bower and John A. Dixon*

A comprehensive guide to the latest techniques for assessing and quantifying the environmental impact of development projects. Much of the book focuses on benefit-cost analysis, but such alternative strategies as input-output analysis and mathematical programming are also discussed, all with an emphasis on practical application.

No other volume covers both natural systems and economic analysis techniques, or integrates the two, as thoroughly as this one. Each chapter is largely self-contained and written in plain, jargon-free language.

**\$25.00** hardcover, **\$10.95** paperback

## **Working for the Sovereign**

EMPLOYEE RELATIONS IN THE FEDERAL GOVERNMENT

*Sar A. Levitan and Alexandra B. Noden*

"A well-written and timely overview of labor-management relations in the federal civil service."—*Ronald W. Haughton, Chairman, Federal Labor Relations Authority*

The first comprehensive study of employee relations and pay structure within the federal government. The authors address contemporary concerns about the civil service in the context of two decades of labor relations, reviewing the structure of federal personnel management, the rise of federal unions, government experiences with collective bargaining, and the machinery used to determine—and justify—the pay, fringe benefits, and job security of federal workers. **\$14.95**

*Policy Studies in Employment and Welfare, no. 39*

**The Johns Hopkins University Press**

Baltimore, Maryland 21218

NEW BOOKS-NEW BOOKS-NEW BOOKS-NEW BOOKS-NEW

V

# Yale's Best in Economics

## **The Financial Development of India, Japan, and the United States**

*A Trilateral Institutional, Statistical, and Analytic Comparison*  
Raymond W. Goldsmith

Goldsmith highlights the essential differences between the financial structures of India and Japan and compares them to that of the United States, regarded as the prototype in this field. The data for India and Japan are taken from two companion volumes published simultaneously (see below) which offer much more detail as well as documented statistical evidence. \$12.95

*Also by Raymond W. Goldsmith*

## **The Financial Development of Japan, 1868-1977**

## **The Financial Development of India, 1860-1977**

These two parallel volumes describe and analyze changes in the financial superstructure and the underlying infrastructure of income and wealth of India and Japan, respectively. Each volume includes over a hundred tables; the books treat in detail the institutional and statistical aspects of the development of capital formation and saving, of financial institutions, of financial instruments, and of the methods of financing agriculture, nonagricultural households, nonfinancial business, and government. Issue ratios for sector and for financial instruments and national balance sheets for a dozen benchmark dates are among the novel features of these valuable studies. Japan volume \$30.00; India volume \$35.00

## **Industrial Innovation in the Soviet Union**

edited by Ronald Amann and Julian Cooper

This sequel to *The Technological Level of Soviet Industry* "is a fascinating compendium about how Soviet industry works (and does not work)...Should be obligatory reading for all." —*The Economist*

"Will clearly become indispensable for anyone wanting to look in sharper focus at the performances of individual Soviet industries." —Michael Simmons, *The Guardian* \$60.00

## **Economic Growth in Prewar Japan**

Takafusa Nakamura

translated by Robert A. Feldman

This is a comprehensive historical survey of Japan's economic performance from the Meiji Restoration to the beginning of the Pacific War. A distinguished Japanese economist, Nakamura has integrated theory, history, and statistical evidence into a balanced presentation accessible to all students of economic history. \$35.00

## **The Rise and Decline of Nations**

*Economic Growth, Stagflation, and Social Rigidities*

Mancur Olson

"Elegant, readable...A convincing little book that could make a big difference in the way we think about modern economic problems." —Peter Passell, *The New York Times Book Review* \$14.95

## **The Political Economy of Growth**

edited by Dennis C. Mueller

This volume represents the first serious testing of Mancur Olson's major new theory on economic performance. \$23.50

**Yale University Press**

**New Haven and London**

## New from LexingtonBooks

### Resources and Energy

#### An Economic Analysis

Ferdinand E. Banks, University of Uppsala

Topics include global views of the economics of natural gas, coal, and uranium, petrochemicals, refining, futures markets, and exploration. 368pp. ISBN 0-669-05203-5 \$34.95

### Innovative Electric Rates

#### Issues in Cost-Benefit Analysis

edited by Sanford V. Berg, University of Florida

Investigates demand forecasting, technological and corporate financial constraints, consumer responses, and regulatory incentives. 352pp. ISBN 0-669-04835-6 \$32.95

### Strategic Planning for Smaller Businesses

#### Improving Corporate Performance and Personal Reward

David A. Curtis, David A. Curtis & Associates

Analyzes the benefits of strategic planning in light of the differences between large and small businesses. 224pp. ISBN 0-669-06011-9 \$21.95

### Current Issues in

#### Public-Utility Economics

Essays in Honor of James C. Bonbright edited by Albert L. Danielson and David R. Kamerschen, The University of Georgia

Expert contributors examine the make-up and functions of state regulatory commissions, the impact of legislation and competition on rate structures, and means of obtaining capital. 352pp. ISBN 0-669-05440-2 \$34.95

### Competitive Structure of the International Banking Industry

Seung H. Kim and Stephen W. Miller, St. Louis University

Foreword by Alfred F. Miossi  
Investigates market segments, financial characteristics, and the legal and regulatory environment. 255pp. ISBN 0-669-05189-6 \$25.95

### Full Employment and Public Policy: The United States and Sweden

Helen Ginsburg, Brooklyn College — City University of New York

Compares the development of full-employment theory in the United States and Sweden, and investigates the social effects of policy decisions. 256pp. ISBN 0-669-01518-8 \$24.95

### Philosophical and Economic Foundations of Capitalism

edited by Svetozar Pejovich, Texas A&M University

Discusses basic institutions of capitalism and socialism and probes the nature of their social processes and their morality. 160pp. ISBN 0-669-05906-4 \$19.95

### Oil-Futures Markets

#### An Introduction

William G. Prast and Howard L. Lax, Atlantis, Inc.

Describes how the international petroleum trade is structured, examines the workings of oil-futures markets in the U.S. and the U.K., and assesses the future of this field. 208pp. ISBN 0-669-06354-1 \$23.95

### If Not for Profit, for What?

A Behavioral Theory of the Nonprofit Sector Based on Entrepreneurship  
Dennis R. Young, State University of New York at Stony Brook and The Program on Non-Profit Organizations, Yale University

Foreword by John G. Simon  
Investigates the effects of changes in public policy on entrepreneurial screening and hence on the behavior of the nonprofit sector. 192pp. ISBN 0-669-06154-9 \$20.95



**LexingtonBooks**

D. C. Heath and Company  
125 Spring Street  
Lexington, MA 02173

(617) 862-6650 (212) 924-6460

**Call our toll-free number**  
800 428-8071

# Military Spending

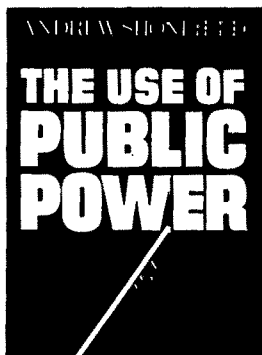
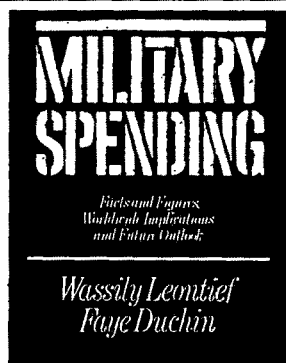
**Facts and Figures, Worldwide  
Implications, and Future Outlook**

**WASSILY LEONTIEF and FAYE DUCHIN**

Using the input-output methodology, the authors look at current trends toward accelerated military spending and assess the impact of some alternatives for fifteen individual regional economies and the world economy as a whole. Their conclusions are striking: they paint a vivid picture of how the military sector drains resources away from other areas of the economy.

They present six scenarios, each describing an alternative development path for the world economy until the year 2000. The authors found that in virtually all regions, despite economic differences, decreased military spending would be accompanied by increases in per capita consumption as well as increases in total world production and trade levels. Cutting back international trade in military goods has a similar effect. Finally, the authors demonstrate that reduced military spending in the poorest, least developed regions supplemented by massive transfers of economic aid from rich, industrialized nations could markedly improve their standard of living by the year 2000.

**1983 240 pp. \$19.95**



## The Use of Public Power

**ANDREW SHONFIELD**

*Edited by ZUZANNA SHONFIELD  
with a Foreword by SIR JOHN HICKS*

Bringing the analysis of *Modern Capitalism* up to the present, *The Use of Public Power* examines the balance between public and private power during the boom years of the sixties and the uncertainties of the

seventies, showing how the persistence of inflation, and popular reaction to it, offered conservative politicians a heaven-sent opportunity of reversing the trend toward increasing public expenditure. Shonfield attacks the monetarist, non-interventionist policies of the Reagan and Thatcher governments and compares the economic achievements of the United States and Britain with those of France, Germany, Japan, and other Western-style mixed economies. He critically appraises the future of the mixed economy.

**1982 160 pp. \$25.00**

*At your bookstore or send your check to Box 900*

**OXFORD UNIVERSITY PRESS**  
**200 Madison Ave. New York, N.Y. 10016**

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## **New in Paperback**

A lively challenge to economists of all persuasions from Marx to Friedman.

"An original and penetrating analysis of the great plague of our times."

—Robert L. Hielbroner

"The economics profession has basically come to the conclusion that wage and price controls are unworkable and uncesirable, [so] it is well worth listening to [Slawson], who thinks that they can be made to work. —Lester C. Thurow

Paper, \$8.95. Cloth, \$16.50

# **W. David Slawson**

# **THE NEW INFLATION**

## **The Collapse of Free Markets**

**Princeton University Press**

41 William Street,  
Princeton, NJ 08540

### **Reasonomics in the Stagflation Economy**

*Sidney Weintraub and Marvin Goodstein, editors*

A distinguished group of economists here assesses the consequences of supply-side and monetarist policies on a wide range of long-term issues and problems. The book treats four main sets of themes: "Supply-Siders, Minorities, and Productivity"; "Monetary and Fiscal Aspects"; "Money, Wages, Controls, and Energy"; and "International Dimensions."

200 pages, \$7.95 paper, \$20.00 cloth

*The first volume in the Post Keynesian Economics series.*

### **Profit Theory and Capitalism**

*Mark Obrinsky*

This book provides an analytical and critical survey of profit in economic theory. Its objectives are to discover the causes of the present, unsatisfactory state of profit theory and to locate the basis for developing a theory that is useful, meaningful, and consistent.

176 pages, \$8.95 paper, \$18.00 cloth

*A new volume in the Post Keynesian Economics series.*

### **New in Paperback**

### **Toward a New U.S. Industrial Policy?**

*Edited, with an Introduction, by  
Michael L. Wachter and Susan M. Wachter*

"If you want one book that covers the breadth of the industrial policy controversy, as it is carried on among thoughtful and sophisticated people, this one will serve you well." —*Washington Post Book World*

536 pages, \$12.95 paper

### **China's Economic Reforms**

*Edited by Lin Wei and Arnold Chao*

This collection presents a comprehensive, up-to-date view of the current economic situation in China from a distinguished group of economists deeply involved with economic policy in the People's Republic.

352 pages, \$25.00 cloth

# **UPP**

**University of Pennsylvania Press**

3933 WALNUT STREET | PHILADELPHIA 19104

# Oxford

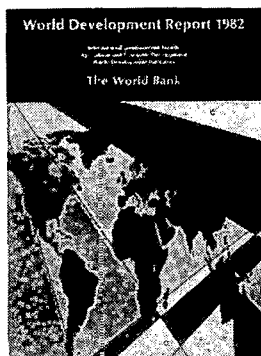
## World Development Report 1982

International Development Trends

Agriculture and Economic  
Development

World Development Indicators

THE WORLD BANK



*World Development Report 1982* opens with emphasis on the more general problems facing developing countries as well as their relations with the developed economies. Part I discusses the low rate of economic growth in industrial nations, high interest rates, and the growing difficulty developing countries have in obtaining concessionary aid. Part II concentrates on agriculture, still the chief source of income for close to two-thirds of the population in developing countries and for the vast majority of the world's poor. The discussion draws on The World Bank's experience in financing some 800 agricultural and rural development projects in more than 70 countries, supplemented by broad, intensive programs of economic, scientific, and social research. As in previous years, the final portion of the *Report*, "World Development Indicators," presents economic and social profiles of more than 120 countries, a highly informative supplement to the wealth of tables, maps, and graphics throughout the *Report*.

1982 182 pp.; charts, tables, graphs cloth \$20.00 paper \$8.00

### Key Features

- Numerous tables present the most up-to-date figures available.
- An abundance of multicolor maps and graphics provides detailed analysis in a clear, attractive format.
- Graphics include long sidebars for close sequential analysis.
- Case studies provide additional useful detail.
- Section on "World Development Indicators" offers 25 two-page tables profiling a comprehensive cross-section of more than 120 countries, with technical notes on the sources and derivation of the data.
- Text includes definitions of economic country-groups as well as a glossary of official acronyms and initials of major development organizations.

### Commentary on earlier editions:

"An authoritative and deeply illuminating summary of worldwide progress against poverty, and of the intricate relationships—not all of them economic—on which it depends."—*The Washington Post*. "Essential reading for any individual or organization interested in or involved with developing countries."—*The Sudan Progress*. "[A] most remarkable publication. It is the nearest thing to having an annual report on the present state of the planet and the people who live on it.... It is going to be the essential almanac for monitoring the way we are going, and where we are going."—*The Guardian* (London)

To order, send payment to:

*Price is subject to change.*

**OXFORD UNIVERSITY PRESS**  
200 Madison Avenue, New York, New York 10016

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## The European Economy

### Growth and Crisis

*Edited by* ANDREA BOLTRO, *Magdalen College, University of Oxford*. This survey of the economic development of Western Europe from the early 1950s to the late 1970s is concerned with two major questions: what forces propelled Europe to unprecedented growth rates in the 1950s and 1960s; and what were the main reasons for the deterioration in performance that took place in the 1970s? Particular attention is devoted to the role of government in these events to demonstrate how economic policies were important in shaping Europe's postwar history and had, on the whole, beneficial effects.

1982 650 pp.; 50 diagrams cloth \$49.00 paper \$19.95

## Economic Foundations of the Gold Problem

*Edited by* ALBERTO QUADRIO-CURZIO, *Catholic University of Milan*. This series of papers was delivered at the 1982 World Conference on Gold by an international group of academic specialists, leading bankers, and practicing producers and dealers. It is probably the most comprehensive survey of the current status of gold in the world economy. The papers are grouped into three sections: Markets and Intermediaries, Medium to Long-term Structural Aspects, and National and Supranational Authorities.

1983 250 pp. \$45.00

## Towards a Political Economy of Urbanization in Third World Countries

*Edited by* HELEN I. SAFA, *Director, Center for Latin American Studies, University of Florida*. The essays in this volume bring a new perspective to bear on the urbanization problem in the Third World, and particularly on the survival mechanisms of the poor. The authors include anthropologists, geographers, economists, and sociologists, who examine the impact of colonialism on the economies of the Third World countries, their dependence on advanced industrial societies for capital and technology, their urban class structures, and the role of their governments in shaping the urban process.

1982 315 pp. \$9.95

## Stalinist Economic Strategy in Practice

### The Case of Albania

ADI SCHNYTZER, *Griffith University, Brisbane, Australia*. The Albanian economy represents a fascinating and unique example of the application of Stalin's economic strategy as set out in his *Economic Problems of Socialism in the USSR*, which was never implemented as policy in the USSR itself. Schnytzer traces the changes in the structure of the Albanian economic planning system between 1945 and 1978, and relates these changes to the pressures of a campaign of forced industrialization. (*Economies of the World*)

1982 180 pp.; 26 tables \$36.00

## Palanpur

### The Economy of an Indian Village

C.J. BLISS, *University of Oxford*, and N.H. STERN, *University of Warwick*. This ambitious and systematic study tests theories of underdevelopment in relation to the motives and behavior of poor farmers in an Indian village. Reporting on the village, its population, and institutions, and providing a review of the development models on which they have drawn, the authors discuss the implications of the research findings for development theory, for policy, and for the future of Palanpur.

1982 340 pp. \$37.50

*Prices and publication dates are subject to change.*

**OXFORD UNIVERSITY PRESS**

200 Madison Avenue, New York, New York 10016

THE COMMISSION  
OF THE EUROPEAN COMMUNITIES  
Presents:

# EUROPEAN ECONOMY

## EUROPEAN ECONOMY

This periodical, which appears four times a year in March, July, September and November, is the main source of information on the Commission's analyses of macroeconomic problems and its proposals for their solution. It gives a review of the current economic situation in the E.C., together with reports and studies on problems of current interest for economic policy. It is accompanied by the following three series of supplements:

**Series A - Recent economic trends** appears monthly, except in August, and describes with the aid of tables and graphs the most recent trends of industrial production, consumer prices, unemployment, the balance of trade, exchange rates and other indicators. It also describes the Commission's macroeconomic forecasts and provides a chronology of economic policy measures in the European Community.

**Series B - Economic prospects: business survey results** reports the main results (orders, stocks, production outlook, etc.) of opinion surveys of industrial chief executives in the E.C. It also appears monthly, with the exception of September.

**Series C - Economic prospects: consumer survey results** reports on the consumer survey carried out three times a year (in January, May and October) throughout the European Community (except Luxembourg) and measures consumers' opinion on the economic situation and outlook.

Cut off and mail to:

**Order Form**  
**OFFICE FOR OFFICIAL PUBLICATION OF THE EUROPEAN COMMUNITIES**  
**L - 2985 Luxembourg**

Please send me No. .... copy(ies) in ..... language of:

	BFR	IRL	UKL	USD
<input type="checkbox"/> European Economy (4 issues per year)	800	13.50	11.60	22.80
<input type="checkbox"/> Series A - Recent economic trends (11 issues per year)	400	6.75	5.80	11.50
<input type="checkbox"/> Series B - Economic prospects - business survey results (11 issues per year)	400	6.75	5.80	11.50
<input type="checkbox"/> Series C - Economic prospects - consumer survey results (3 issues per year)	150	2.50	2.20	4.20
<input type="checkbox"/> All three supplements	950	10.00	13.80	27.00
<input type="checkbox"/> Combined subscription: European Economy and supplements	1,750	29.40	25.50	50.00

Prices exclude VAT in Luxembourg.

Payment is due on receipt of invoice.

Name: ..... Address: ..... AER

..... Date: ..... Signature: .....

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

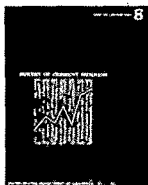
# FIVE ESSENTIAL TOOLS FOR ECONOMISTS

Five monthly periodicals  
about economic data  
published by the Federal  
agencies responsible  
for collecting and  
interpreting the data



## MONTHLY LABOR REVIEW

Current data and analysis on employment, unemployment, prices, wages, productivity, industrial relations, economic growth, foreign labor developments, and job safety. Published by the Bureau of Labor Statistics, U.S. Department of Labor.  
**\$26 per year.**



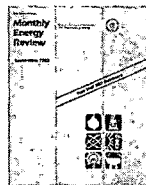
## SURVEY OF CURRENT BUSINESS

Estimates of national, regional, and international economic accounts; articles on the business and economic situation; and a statistical section covering all aspects of the economy. Published by the Bureau of Economic Analysis, U.S. Department of Commerce.  
**\$30 per year.**



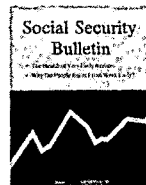
## AGRICULTURAL OUTLOOK

Current analysis and forecast data on the U.S. food and fiber economy, covering commodity supply and demand, farm income, world agriculture and trade, food prices and marketing, farm inputs, transportation, storage, and the general economy. Published by the Economic Research Service, U.S. Department of Agriculture.  
**\$31 per year.**



## MONTHLY ENERGY REVIEW

Current and historical energy statistics for production, consumption, imports, exports, storage, and costs of the major energy resources, including petroleum, natural gas, coal, and electric power. Published by the Energy Information Administration.  
**\$36 per year.**



## SOCIAL SECURITY BULLETIN

Analytical articles and current statistics on Old-Age, Survivors and Disability Insurance, Supplemental Security Income, and Aid to Families with Dependent Children programs. Published by the Social Security Administration, U.S. Department of Health and Human Services.  
**\$29 per year.**

## Order Form

Mail To: Dept 36AD, Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402

Enclosed is \$ \_\_\_\_\_ ☐ check,  
☐ money order, or charge to my  
Deposit Account No.

\_\_\_\_\_-\_\_\_\_

Order No. \_\_\_\_\_

**MasterCard and  
VISA accepted.**



### Credit Card Orders Only

Total charges \$ \_\_\_\_\_ Fill in the boxes below.

Credit Card No. \_\_\_\_\_

Expiration Date  
Month/Year \_\_\_\_\_

Please enter my subscription(s) as follows: ☐ Monthly Labor Review (MLR) \$26  
☐ Survey of Current Business (SCB) \$30 ☐ Agricultural Outlook (AO) \$31  
☐ Monthly Energy Review (MER) \$36 ☐ Social Security Bulletin (SSB) \$29

COMPANY OR PERSONAL NAME

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

ADDITIONAL ADDRESS/ATTENTION LINE

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

STREET ADDRESS

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

CITY

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

STATE

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

ZIP CODE

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

(OR) COUNTRY

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

PLEASE PRINT OR TYPE

### For Office Use Only

Quantity	Charges
_____	Publications _____
_____	Subscription _____
_____	Special Shipping Charges _____
_____	International Handling _____
_____	Special Charges _____
_____	OPNR _____
_____	UPNS _____
_____	Balance Due _____
_____	Discount _____
_____	Refund _____

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## **Business Cycles, Inflation, and Forecasting Second Edition**

**Geoffrey H. Moore**

Incorporating current research and up-to-date essays, the new edition of this highly acclaimed book will help decisionmakers evaluate economic statistics and put them to practical use.

June      \$42.50t cloth      \$19.50t paper

## **Exchange Rate and Trade Instability Causes, Consequences, and Remedies**

Edited by **David Bigman** and **Teizo Taya**

Economists from throughout the world shed light on the concern for maintaining national stability in a world economy no single country can hope to control. Focuses on the causes, implications, and remedies of the current period of volatile money markets.

May      \$35.00t

## **Sources of European Economic Information Fourth Edition**

Edited by **Euan Blauvelt** and  
**Jennifer Durlacher**

The major tool for locating essential economic information published in Western Europe. Now expanded to include Eastern Bloc countries, this edition lists and analyzes more than 6000 sources, each indexed by country, subject, and original publisher.

March      \$95.00

## **Stalemate in Technology Innovations Overcome the Depression**

**Gerhard Mensch**

" . . . In many ways the most important contribution to macroeconomic theory since John Maynard Keynes." *World Press Review*

Mensch predicts that we are on the threshold of another burst of technological innovation, and that the nations best able to anticipate the growth areas and capitalize on the markets will emerge as world leaders.

May      \$11.95

## **Steel Upheaval in a Basic Industry**

**Donald F. Barnett** and **Louis Schorsch**

The decline of the American steel industry over the past thirty years presented as an illuminating case study of the overall decline of basic U.S. industries.

June      \$28.00t

## *From the Peace Science Studies Series*

## **Conflict Analysis and Practical Conflict Management Procedures An Introduction to Peace Science**

**Walter Isard** and **Christine Smith**

February      \$35.00

## **International and Regional Conflict**

### **Analytic Approaches**

Edited by **Walter Isard** and **Yoshimi Nagao**

June      \$36.00t

# **Ballinger**

**PUBLISHING COMPANY** • A Subsidiary of Harper & Row, Publishers, Inc.  
54 Church Street • Cambridge, MA 02138

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# Profit from the best in economics.

## **ECONOMICS, 3rd Edition, 1983**

William P. Albrecht, Jr., University of Iowa  
1983 784 pp. (22434-5)

## **ECONOMICS, 1983**

W. H. Locke Anderson, Ann F. Putallaz,  
and William G. Shepherd, all of the  
University of Michigan  
1983 912 pp. (22429-5)

—also available in paperback:

**Microeconomics (58103-3);**

**Macroeconomics (54281-1)**

## **MONEY, THE FINANCIAL SYSTEM, AND MONETARY POLICY, Second Edition, 1983**

Thomas F. Cargill, University of Nevada,  
Reno  
1983 608 pp. (60036-1)

## **MANAGERIAL ECONOMICS: Theory, Practice, and Problems, Second Edition, 1983**

Evan J. Douglas, Concordia University,  
Montreal, Canada  
1983 576 pp. (55021-0)

## **MACROECONOMICS: Theory and Policy, Second Edition, 1983**

Michael R. Edgmand, Oklahoma  
State University  
1983 464 pp. (54268-8)

## **INTERNATIONAL ECONOMIC RELATIONS, 1983**

John S. Hodgson and Mark G. Herander,  
both of the University of South Florida  
1983 590 pp. (47275-3)

## **MONETARY POLICY AND THE FINANCIAL SYSTEM, Fifth Edition, 1983**

Paul M. Horvitz, University of Houston, and  
Richard A. Ward, University of Southern  
California, Los Angeles  
1983 576 pp. (59993-6)

## **ANTITRUST AND TRADE REGULATION: Selected Issues and Case Studies, 1983**

Marshall C. Howard, University of  
Massachusetts, Amherst  
1983 288 pp. (03834-9)

## **PRINCIPLES OF PUBLIC FINANCE, 1983**

Charles W. Meyer, Iowa State University and  
J. Ronnie Davis, Western Washington  
University  
1983 448 pp. (70988-1)

## **INTERNATIONAL TRADE AND FINANCE: Theory and Policy, 1983**

Holley Ulbrich, Clemson University  
1983 448 pp. (47395-9)

For further information, or to order or reserve  
examination copies, please write: Robert Jordan,  
Dept. CJ-215, Prentice-Hall, Inc., Englewood Cliffs,  
NJ 07632

For "super-quick" service, dial  
TOLL-FREE (800) 526-0485\* between  
9:00 a.m. and 4:30 p.m., (EST).

\*not applicable in New Jersey, Alaska, Hawaii, or Puerto Rico.



# Prentice-Hall

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# The GENEVA PAPERS on Risk and Insurance

A quarterly Journal on Risk, Uncertainty and Insurance Economics

## The theory of Risk and Uncertainty:

### Major topics:

- Kenneth Arrow, Jacques Drèze, Martin Feldstein, Karl Borch, Joseph Stiglitz and Raymond Barre have contributed to the yearly "Annual Lecture of the Geneva Association" special issues;
- one issue is devoted every year to selected papers and essays from academic research.

## The practice of Risk and Uncertainty management:

- studies carried out for The Geneva Association on specific industrial areas (the future risks of computers' utilisation, the vulnerability of containers, the consequential losses in the chemical industry, recall practices in industry), by reputed research centres (SRI, Battelle, Diebold, Prognos, Technomic);
- studies promoted by the Geneva Association in the field of finance, savings, insurance services, simulation models.

### CONTENTS OF VOLUME 7, 1982:

#### No 22 STUDIES IN RISK MANAGEMENT (I)

Risk Management Under Changing Economic Conditions, by Robert L. Carter; Développement Économique et Croissance des Risques, par Orlin Glarini; Recall Practices Among Manufacturers of Consumer Products, by Roy Damary and A.S. Hurst.

#### No 23 STUDIES IN RISK MANAGEMENT (II)

The Risks from Liquefied Natural Gas, by Emilio Blomonte; A Comparison of Attitudes Towards Risks Among Business Managers, by Gordon C.A. Dickson; The Composition of a Consequential Loss, by Brian J. Kylen; An Overview of Risk Management, by Douglas G. Olson, John A. Simkins, Jr.; The Captive Insurance Phenomenon, by H. Felix Kloman, D. Hugh Rosenbaum; Managing Insurance Risks, by Maurice Salvator; The Bibliography and History of Risk Management, by G. Neff Crockford; The Risk Management Function and Education in the United States, by George L. Head.

#### No 24 ESSAYS IN INSURANCE ECONOMICS

Optimum Retirement Age, by Norma L. Larsen and August Reistad; Demand for Supplementary Health Insurance in Switzerland: A Theoretical and Empirical Investigation, by Peter Zweifel; Business Insurance and Large Corporations, by Brian G.M. Main; Some Considerations on Goal Systems of Insurance Companies, by Bernd Kaluzs; Indexed and Non-Indexed Insurance, by Yaffa Mechren.

#### No 25 THE ASIR MODEL (The Advanced Simulation Model of Insurance and Re-insurance Operations)

Inflation and Interest Rates: A Study Using the ASIR Model, by Margaret Brown; Fluctuating Exchange Rates: A study Using the ASIR Model, by Lawrence Galtz; From the User's Manual: 1. Main Design Concepts, 2. Preparing Data for a Simulation, 3. Running the ASIR Model, 4. Data Requirements for ASIR.

The Subscription Price for one volume (4 issues) is 80 Swiss Francs (40 U.S. dollars). The price of each issue is 25 Swiss Francs (12 U.S. dollars approximately).

Distribution: Librairie Georg & Cie SA, Corratierie 21  
1211 Geneva 11 (Switzerland)

SPECIMEN COPIES ARE AVAILABLE FROM  
THE GENEVA ASSOCIATION (18, chemin Rieu, CH-1208 Geneva)

Please mail this coupon to:  
The Geneva Association  
18, chemin Rieu  
1208 Geneva (Switzerland)

Please send me a free inspection copy of  
THE GENEVA PAPERS

Name: .....

Adresse: .....

NOW IN VINTAGE PAPERBACK

**"The best book yet about  
what industrial  
policy might  
be in the  
United States"**

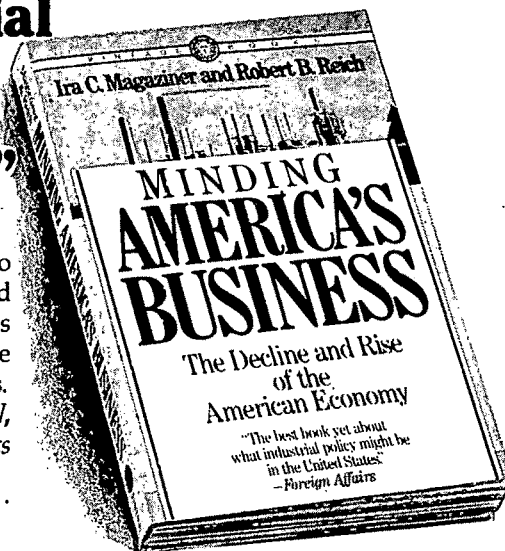
—Foreign Affairs

"An excellent introduction to  
explaining why the United  
States needs industrial policies  
and what other countries are  
doing in their industrial policies.

—LESTER THUROW,  
*New York Review of Books*

"Extraordinarily detailed...  
lucid and often intriguing."

—ROBERT J. SAMUELSON,  
*New Republic*



Illustrated with charts and graphs. 400 pages.  
\$5.95, now at your bookstore



**VINTAGE BOOKS**  
A division of Random House

**PUBLICATIONS FROM THE  
UNITED NATIONS**

**WORLD ECONOMIC SURVEY 1981-1982**

*Current trends in the World Economy.* Gives the growth in  
the world economy and current policy stances, interna-  
tional trade and payments, adjustment and international  
capital flows to developing countries.

E.82.II.C.1

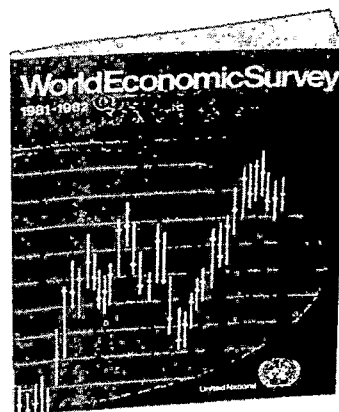
\$9.00

**TOWARDS THE NEW INTERNATIONAL  
ECONOMIC ORDER**

*Report of the Director-General for Development and  
International Economic Co-operation.*

E.82.II.A.7

\$9.00



**UNITED NATIONS**

Room A-3315  
New York, N.Y. 10017



**PUBLICATIONS**

Palais des Nations  
1211 Geneva 10, Switzerland

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# **AEA sponsored Group Life Insurance for you and your family— at attractive rates!**

The AEA Group Life Insurance Plan can help provide valuable supplementary protection—at attractive rates—for eligible members and their dependents.

Because AEA has joined a large Insurance Trust which includes other scientific and technical organizations, the low cost may be even further reduced by dividend credits. In the past seven years, insured members received credits on their April 1 semiannual payment notices averaging over 45% of their annual premium contributions. (These credits are based on the amount paid during the previous policy year ending September 30.) Of course future dividends and credits, and their amounts, cannot be promised or guaranteed.

Now may be a good time for you to re-evaluate your present coverage and look into AEA Life Insurance. Just fill out and return the coupon for more details at no obligation.

**Administrator, AEA Group Insurance Program**  
1707 L Street, N.W.—Suite 700  
Washington, D.C. 20036

E - 1

Please send me more information about the AEA Life Insurance Plan.

Name \_\_\_\_\_ Age \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Or—call today Toll-Free 800-424-9883  
(Washington, DC area, call 296-8030)

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# INDEX OF ECONOMIC ARTICLES

prepared under the auspices of

*The Journal of Economic Literature*  
of the  
*American Economic Association*

- ✓ Each volume in the **Index** lists articles in major economic journals and in collective volumes published during a specific year.
- ✓ Most of the **Index's** volumes also include articles of testimony from selected congressional hearings in government documents published during the year.
- ✓ No other single reference source covers as many articles classified in economic categories as the **Index**.
- ✓ The 1977 volume contains over 10,500 entries.

## Currently available are:

Volume	Year Covered
XI	1969
XII	1970
XIII	1971
XIV	1972
XV	1973
XVI	1974
XVII	1975
XVIII	1976
XIX	1977
XX	1978 (in preparation)

*an  
indispensible  
tool for...*

**ECONOMISTS  
REFERENCE LIBRARIANS  
RESEARCHERS  
TEACHERS  
STUDENTS  
AUTHORS**

*Future volumes will be published regularly  
to keep the series as current as possible.*

**Price: \$50.00 per volume (special 30% discount to  
AEA members)**

Distributed by:

**RICHARD D. IRWIN, INC.** Homewood, Illinois 60430

# ANNOUNCING



Sixth Edition

## **Guide to Graduate Study in Economics and Agricultural Economics**

in the United States of America and Canada

Designed to provide students anticipating graduate study in economics and agricultural economics, and their advisors, with information on available graduate training programs.

Includes descriptions of 262 graduate programs, supplemented by comparative data and information for prospective students, domestic and foreign.

**PUBLISHED BY THE ECONOMICS INSTITUTE,**

University of Colorado at Boulder, Boulder, Colorado 80309 under the auspices of the **American Economic Association** and the **American Agricultural Economics Association.**

PRICE: \$19.95 per copy

### ORDER FORM

RICHARD D. IRWIN, INC.  
1818 Ridge Road  
Homewood, Illinois 60430

Please send me \_\_\_\_\_ copies of *Guide to Graduate Study in Economics and Agricultural Economics in the United States of America and Canada*, 6th edition.

Enclosed is my check/money order for \$\_\_\_\_\_ (\$19.95 per copy).

Please print or type:

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_